



(12)发明专利

(10)授权公告号 CN 104503965 B

(45)授权公告日 2017.08.29

(21)申请号 201410548447.2

(56)对比文件

(22)申请日 2014.10.16

CN 102521389 A, 2012.06.27,

(65)同一申请的已公布的文献号

CN 103049579 A, 2013.04.17,

申请公布号 CN 104503965 A

US 2013238656 A1, 2013.09.12,

(43)申请公布日 2015.04.08

纪红波.PostgreSQL数据库集群基本技术分析与实现.《吉林工商学院学报》.2010,第26卷(第5期),第69-72页.

(73)专利权人 杭州斯凯网络科技有限公司

审查员 胡璇

地址 310013 浙江省杭州市紫荆花路2号联合大厦B座10楼

(72)发明人 周正中

(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

(51)Int.Cl.

G06F 17/30(2006.01)

权利要求书3页 说明书8页 附图7页

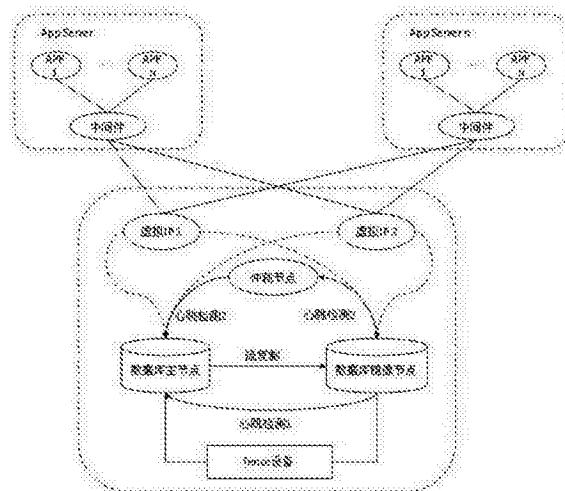
H04L 29/08(2006.01)

(54)发明名称

PostgreSQL高弹性的高可用及负载均衡实现方法

(57)摘要

本发明公开了一种PostgreSQL高弹性的高可用及负载均衡实现方法。属于数据库技术领域，该方法实现了数据库层面完全无单点故障的风险，并且在连接层面使用会话保持来解决应用感知的问题，还实现了对应用完全透明的高可用 failover。连接池和应用程序建立的TCP连接不受failover影响，因此这个会话不会中断。连接池和数据库的连接断开后自动连接。业务层在切换期间的会话自动回滚。在连接池或驱动层配置好多个对应的后端数据库连接，根据配置的算法，数据插入的SQL自动分发到后端数据库以实现负载均衡。当后端数据库无响应时，自动分发到下一个候选连接，以实现failover。



1. PostgreSQL高弹性的高可用及负载均衡实现方法,其特征在于,包括:读写混合场景实现方法和写场景实现方法;

(1) 读写混合场景实现方法步骤如下:

(1.1) 准备3台服务器,其中2台服务器分别用于数据库主节点和数据库镜像节点,另1台服务器用于仲裁节点;

(1.2) 分别将数据库主节点服务器的主机、数据库镜像节点服务器的主机和仲裁节点服务器的主机配置在同一个网段,并设数据库主节点为node1,设数据库镜像节点为node2,设数据库的一个虚拟IP为虚拟IP1,设数据库的另一个虚拟IP为虚拟IP2,设仲裁节点为VOTE_HOST;

(1.3) 配置数据库主节点和数据库镜像节点的FENCE设备,为了防止发生脑裂现象,让FENCE设备在自动failover前关闭服务器;

(1.4) 配置数据库主节点和数据库镜像节点的网络,编写虚拟IP配置文件;

(1.5) 在数据库主节点和数据库镜像节点安装PostgreSQL数据库软件;

(1.6) 在数据库主节点初始化数据库;

(1.7) 在数据库主节点配置数据库访问控制配置文件pg_hba.conf、postgresql.conf和recovery.conf,这三个数据库访问控制配置文件都用于数据库复制、启动数据库、创建流复制用户和启动虚拟IP;

(1.8) 创建镜像数据库,配置数据库镜像节点流复制环境,并启动镜像虚拟IP;

(1.9) 配置仲裁节点,设置用于仲裁网络异常的数据库监听端口,即配置仲裁机,在仲裁机上起一个监听端口,用于跳转到虚拟IP的PostgreSQL主库监听端口;

(1.10) 在数据库主节点创建心跳检查函数,数据库主节点的变更将自动复制到数据库镜像节点;

(1.11) 编写心跳检查软件和仲裁节点检查软件,并把心跳检查软件部署在数据库主节点和数据库镜像节点,把仲裁节点检查软件部署在仲裁节点,用于自动failover;

(1.12) 开启用于自动failover的心跳检查软件;

(1.13) 配置组件监控检查,所述组件包括用于负责心跳检测和故障时激活镜像以及切换虚拟IP的HA切换软件,所述组件还包括用于读写负载均衡和客户端会话保持连接池的中间件,所述组件还包括用于数据库主节点IP读写连接和数据库镜像节点IP读连接的虚拟IP,所述组件还包括用于投票解决单边网络故障的仲裁节点,所述组件还包括用于解决脑裂问题的FENCE设备,所述组件还包括用于负责读写的数据库主节点和从数据库主节点获取实时增量同步且只负责读的数据库镜像节点;

所述HA切换软件的实现过程是:

(1.13.1) 启动,判断主节点是否正常;

(1.13.2) 如果主节点不正常则结束,如果主节点正常则判断本地节点角色是否为镜像节点;

(1.13.3) 如果本地节点不是镜像节点则结束,如果本地节点为镜像节点则计数清零,并检查主节点状态;

(1.13.4) 如果主节点状态健康则结束,如果主节点状态不健康则检查仲裁节点状态;

(1.13.5) 如果仲裁节点状态不健康则再次进行计数清零,如果仲裁节点状态健康则从

仲裁节点检查主节点状态；

(1.13.6) 如果从仲裁节点检查主节点状态为健康状态则再次进行计数清零,如果从仲裁节点检查主节点状态为不健康状态则进行主节点累计异常次数自增计数,然后进行主节点连接异常次数判断1;

(1.13.7) 如果主节点连接异常次数判断1是第一次时则检查镜像节点延迟是否正常,如果主节点连接异常次数判断1不是第一次时则进行主节点连接异常次数判断2;

(1.13.8) 如果检查镜像节点延迟为不正常则再次进行计数清零,如果检查镜像节点延迟为正常则生成镜像节点健康状态标记,然后则进行主节点连接异常次数判断2;

(1.13.9) 如果主节点连接异常次数判断2的异常次数未达到阈值则再次进行计数清零,如果主节点连接异常次数判断2的异常次数达到阈值则检查镜像节点状态标记;

(1.13.10) 如果检查镜像节点状态标记为正常则激活镜像、切换虚拟IP和生成切换标记,并在激活镜像、切换虚拟IP和生成切换标记后结束;如果检查镜像节点状态标记为不正常则生成主节点不健康标签和不发生切换,在生成主节点不健康标签和不发生切换后结束;

(1.14) 配置用于读写分离的连接池;所述读写分离的实现过程是:APP提交SQL来判断SQL是读还是写,如果判断SQL是写则分发给主节点,如果判断SQL是读则分发给镜像节点;

(1.15) 人为的switchover测试,关闭数据库主节点,关闭虚拟IP1,激活数据库镜像节点数据库,数据库镜像节点数据库recovery.conf自动修改为recovery.done,切换为主角色,在数据库镜像节点启动虚拟IP1,数据库主节点recovery.done改为recover.conf,启动数据库主节点,角色切换为镜像角色;

(1.16) 通过拔数据库主节点网线,或者关闭数据库主节点,或者关闭数据库主节点服务器,或者关闭数据库主节点网卡来自动进行failover测试;

(1.17) 当failover发生时,数据库镜像节点自动切换为数据库主节点,同时虚拟IP1自动在数据库镜像节点开启;

(2) 写场景实现方法步骤如下:

(2.1) 准备至少2台用于安装数据库的主机;

(2.2) 在每台主机上安装数据库软件;

(2.3) 在每台主机执行初始化数据库操作,并配置数据库监听端口;

(2.4) 在每台主机创建同样的SCHEMA、用户、密码;

(2.5) 在每台主机创建同样的业务表和约束;

(2.6) 配置DNS,使主机名对应多个数据库主机的IP,同时配置监控,当数据库节点发生变更时实时更新DNS条目;

(2.7) 配置连接池或驱动来实现写负载均衡以及HA;所述写负载均衡以及HA的实现过程是:APP提交SQL,由SQL分发策略采用写的方式分发给节点1,分发给节点1后判断分发是否成功,如果分发成功则分发完成,如果分发不成功则SQL分发策略尝试分发给下一个节点,直到分发成功为止则实现分发完成作业;

(2.8) 应用软件通过pgbouncer连接数据库,DNS策略配置为round-robin模式,当新建连接时,轮询的选择DNS解析出来的IP,从而POOLSERVER分布在不同的后端数据库,实现负载均衡。

2. 根据权利要求1所述的PostgreSQL高弹性的高可用及负载均衡实现方法,其特征在于,在步骤(1.3)中,为避免FENCE慢或者FENCE不成功,还需要关闭数据库服务器操作系统的acpi服务,同时开启idrac的ipmi功能,并给用户赋予ipmi可开关机的OPERATOR角色。

3. 根据权利要求1所述的PostgreSQL高弹性的高可用及负载均衡实现方法,其特征在于,在步骤(1.7)中,在数据库主节点和数据库镜像节点配置流复制密码文件;在数据库主节点启动数据库,添加replication数据库角色;在数据库主节点启动虚拟IP;在数据库主节点配置数据库访问控制文件pg_hba.conf,为了在集群脚本中要用到更新sky_pg_cluster数据库的表,则需要允许主节点、镜像节点、虚拟IP和仲裁节点以及回环地址通过sky_pg_cluster用户访问sky_pg_cluster数据库。

4. 根据权利要求1所述的PostgreSQL高弹性的高可用及负载均衡实现方法,其特征在于,在步骤(1.10)中,插入初始数据,创建测试函数,用于测试数据库是否正常,包括所有表空间的测试,使用update不同的表空间中的数据,并不能立刻反映表空间的问题;因为大多数数据在shared_buffer中,如果表空间对应的文件系统io有问题,那么在checkpoint时会产生58类的错误,使用pg_stat_file函数来立刻暴露io的问题。

5. 根据权利要求1所述的PostgreSQL高弹性的高可用及负载均衡实现方法,其特征在于,在步骤(1.11)中,配置心跳需要的密码文件,集群failover软件的sky_pg_clusterd.sh将用这个密码文件分别用于访问虚拟IP上的PostgreSQL监听端口、本机standby的PostgreSQL监听端口和VOTE_HOST上的跳转端口。

PostgreSQL高弹性的高可用及负载均衡实现方法

技术领域

[0001] 本发明涉及数据库技术领域,尤其涉及一种PostgreSQL高弹性的高可用及负载均衡实现方法。

背景技术

[0002] 在大型的业务系统中,数据库一般处于比较核心的地位,例如涉及用户信息,用户账户信息,用户行为信息的存储。用户或用户之间信息的交互都需要数据库的支持,数据库故障将导致核心业务系统故障。数据库的不间断运行成为业务系统稳定性关键因素。

[0003] 传统数据库高可用方法有两种应用场景:

[0004] 第一种应用场景是利用存储设备的复制功能来实现高可用。参见图6所示,把一个数据库存储设备中的数据采用存储复制的方式复制到另一个数据库存储设备中去,App1或者Appn访问虚拟IP,通过failover决定虚拟IP是通过数据库活动实例1来访问一个数据库,还是通过数据库非活动实例n来访问另一个数据库。这种方法的缺陷是:(1)数据库的特征受制于硬件存储设备,需要存储硬件厂商结合数据库的特征设计,无法支持所有数据库,在没有经过厂商认证的数据库品牌冒然使用,可能导致数据一致性问题或数据块损坏问题;(2)成本高,存在高昂的软件许可成本和硬件成本;(3)存储层面的同步无法和应用结合,无法实现同步或异步的同步,如果是同步复制,那么将增加故障点,同时带来性能损失;如果是异步复制则会增加丢失数据的风险。

[0005] 第二种应用场景是利用共享存储和高可用软件实现,参见图7所示,App1或者Appn访问虚拟IP,通过failover决定虚拟IP是通过数据库活动实例1来访问数据库共享存储中的数据,还是通过数据库非活动实例n来访问数据库共享存储中的数据。这种方法基本上适用于所有的数据库产品,这种方法有几个缺陷:(1)成本高,需要支付高昂的存储硬件和高可用软件费用;(2)依赖存储设备的高可用,如果存储故障则使得数据库高可用失效,存在单点故障;(3)数据库主机切换时会中断网络层会话,无法实现应用无感知。

发明内容

[0006] 本发明是为了解决现有数据库在数据库层存在单点故障风险,在连接层面存在应用感知的问题,对failover的应用不透明的这些不足,提供一种PostgreSQL高弹性的高可用及负载均衡实现方法,该方法实现了数据库层面完全无单点故障的风险.并且在连接层面使用会话保持来解决应用感知的问题,还实现了对应用完全透明的高可用failover。

[0007] 为了实现上述目的,本发明采用以下技术方案:

[0008] PostgreSQL高弹性的高可用及负载均衡实现方法,包括:读写混合场景实现方法和写场景实现方法;

[0009] (1) 读写混合场景实现方法步骤如下:

[0010] (1.1) 准备3台服务器,其中2台服务器分别用于数据库主节点和数据库镜像节点,另1台服务器用于仲裁节点;

- [0011] (1.2) 分别将数据库主节点服务器的主机、数据库镜像节点服务器的主机和仲裁节点服务器的主机配置在同一个网段，并设数据库主节点为node1，设数据库镜像节点为node2，设数据库的一个虚拟IP为虚拟IP1，设数据库的另一个虚拟IP为虚拟IP2，设仲裁节点为VOTE_HOST；
- [0012] (1.3) 配置数据库主节点和数据库镜像节点的FENCE设备，为了防止发生脑裂现象，让FENCE设备在自动failover前关闭服务器；
- [0013] (1.4) 配置数据库主节点和数据库镜像节点的网络，编写虚拟IP配置文件；还关闭network服务的自动启动节点node1和node2；并增加该自动启动节点node1和node2的network服务启动项到rc.local。
- [0014] (1.5) 在数据库主节点和数据库镜像节点安装PostgreSQL数据库软件；
- [0015] (1.6) 在数据库主节点初始化数据库；
- [0016] (1.7) 在数据库主节点配置数据库访问控制配置文件pg_hba.conf、postgresql.conf和recovery.conf，这三个数据库访问控制配置文件都用于数据库复制、启动数据库、创建流复制用户和启动虚拟IP；
- [0017] (1.8) 创建镜像数据库，配置数据库镜像节点流复制环境，并启动镜像虚拟IP；
- [0018] (1.9) 配置仲裁节点，设置用于仲裁网络异常的数据库监听端口，即配置仲裁机，在仲裁机上起一个监听端口，用于跳转到虚拟IP的PostgreSQL主库监听端口；
- [0019] (1.10) 在数据库主节点创建心跳检查函数，数据库主节点的变更将自动复制到数据库镜像节点；
- [0020] (1.11) 编写心跳检查软件和仲裁节点检查软件，并把心跳检查软件部署在数据库主节点和数据库镜像节点，把仲裁节点检查软件部署在仲裁节点，用于自动failover；
- [0021] (1.12) 开启用于自动failover的心跳检查软件；
- [0022] (1.13) 配置组件监控检查，所述组件包括用于负责心跳检测和故障时激活镜像以及切换虚拟IP的HA切换软件，所述组件还包括用于读写负载均衡和客户端会话保持连接池的中间件，所述组件还包括用于数据库主节点IP读写连接和数据库镜像节点IP读连接的虚拟IP，所述组件还包括用于投票解决单边网络故障的仲裁节点，所述组件还包括用于解决脑裂问题的FENCE设备，所述组件还包括用于负责读写的数据库主节点和从数据库主节点获取实时增量同步且只负责读的数据库镜像节点；
- [0023] 所述HA切换软件的实现过程是：
- [0024] (1.13.1) 启动，判断主节点是否正常；
- [0025] (1.13.2) 如果主节点不正常则结束，如果主节点正常则判断本地节点角色是否为镜像节点；
- [0026] (1.13.3) 如果本地节点不是镜像节点则结束，如果本地节点为镜像节点则计数清零，并检查主节点状态；
- [0027] (1.13.4) 如果主节点状态健康则结束，如果主节点状态不健康则检查仲裁节点状态；
- [0028] (1.13.5) 如果仲裁节点状态不健康则再次进行计数清零，如果仲裁节点状态健康则从仲裁节点检查主节点状态；
- [0029] (1.13.6) 如果从仲裁节点检查主节点状态为健康状态则再次进行计数清零，如果

从仲裁节点检查主节点状态为不健康状态则进行主节点累计异常次数自增计数,然后进行主节点连接异常次数判断1;

[0030] (1.13.7) 如果主节点连接异常次数判断1是第一次时则检查镜像节点延迟是否正常,如果主节点连接异常次数判断1不是第一次时则进行主节点连接异常次数判断2;

[0031] (1.13.8) 如果检查镜像节点延迟为不正常则再次进行计数清零,如果检查镜像节点延迟为正常则生成镜像节点健康状态标记,然后则进行主节点连接异常次数判断2;

[0032] (1.13.9) 如果主节点连接异常次数判断2的异常次数未达到阈值则再次进行计数清零,如果主节点连接异常次数判断2的异常次数达到阈值则检查镜像节点状态标记;

[0033] (1.13.10) 如果检查镜像节点状态标记为正常则激活镜像、切换虚拟IP和生成切换标记,并在激活镜像、切换虚拟IP和生成切换标记后结束;如果检查镜像节点状态标记为不正常则生成主节点不健康标签和不发生切换,在生成主节点不健康标签和不发生切换后结束;

[0034] (1.14) 配置用于读写分离的连接池;所述读写分离的实现过程是:APP提交SQL来判断SQL是读还是写,如果判断SQL是写则分发给主节点,如果判断SQL是读则分发给镜像节点;

[0035] (1.15) 人为的switchover测试,关闭数据库主节点,关闭虚拟IP1,激活数据库镜像节点数据库,数据库镜像节点数据库recovery.conf自动修改为recovery.done,切换为主角色,在数据库镜像节点启动虚拟IP1,数据库主节点recovery.done改为recover.conf,启动数据库主节点,角色切换为镜像角色;

[0036] (1.16) 通过拔数据库主节点网线,或者关闭数据库主节点,或者关闭数据库主节点服务器,或者关闭数据库主节点网卡来自动进行failover测试;

[0037] (1.17) 当failover发生时,数据库镜像节点自动切换为数据库主节点,同时虚拟IP1自动在数据库镜像节点开启;

[0038] (2) 写场景实现方法步骤如下:

[0039] (2.1) 准备至少2台用于安装数据库的主机;

[0040] (2.2) 在每台主机上安装数据库软件;

[0041] (2.3) 在每台主机执行初始化数据库操作,并配置数据库监听端口;

[0042] (2.4) 在每台主机创建同样的SCHEMA、用户、密码;

[0043] (2.5) 在每台主机创建同样的业务表和约束;

[0044] (2.6) 配置DNS,使主机名对应多个数据库主机的IP,同时配置监控,当数据库节点发生变更时实时更新DNS条目;

[0045] (2.7) 配置连接池或驱动来实现写负载均衡以及HA;所述写负载均衡以及HA的实现过程是:APP提交SQL,由SQL分发策略采用写的方式分发给节点1,分发给节点1后判断分发是否成功,如果分发成功则分发完成,如果分发不成功则SQL分发策略尝试分发给下一个节点,直到分发成功为止则实现分发完成作业;

[0046] (2.8) 应用软件通过pgbouncer连接数据库,DNS策略配置为round-robin模式,当新建连接时,轮询的选择DNS解析出来的IP,从而POOLSERVER分布在不同的后端数据库,实现负载均衡。

[0047] 本方案对于读写混合场景是,利用PostgreSQL数据库实时流复制实现双份数据库

镜像,解决依赖存储的高可用的问题。自定义的数据库心跳和切换逻辑实现数据库 failover,解决购买商业高可用软件的问题。使用仲裁和fence设备解决脑分裂问题。使用连接池会话层保持解决应用感知问题。连接池可部署多个,不存在单点故障。同时连接池可实现读写分离。同步和异步完全由事务决定,程序如果发起同步事务,则事务必须到达镜像节点后才提交完成;如果是异步事务,则事务到达主节点后就可以提交;解决了重要事务绝对不丢失,不重要事务选择异步则减少性能损失;实现了弹性控制。对镜像节点的延迟监控,如果发现镜像延迟达到阈值,告警。为防止数据丢失,在镜像延迟超过阈值时不会发生切换。本方案对于写场景是,记录用户行为数据;使用完全独立的两套硬件,创建相同的 SCHEMA,实现写负载均衡;在连接池或驱动层实现高可用,负载均衡,解决了连接层单点。当主库数据库本身或数据库所在的硬件,或者存储故障,或者存储空间不足等问题发生时,数据库心跳检测程序将检测到问题,使用fence设备截断主库与外界的联系,然后激活主库镜像并切换IP。连接池和应用程序建立的TCP连接不受failover影响,因此这个会话不会中断。连接池和数据库的连接断开后自动连接。业务层在切换期间的会话自动回滚。在连接池或驱动层配置好多个对应的后端数据库连接,根据配置的算法,数据插入的SQL自动分发到后端数据库以实现负载均衡。当后端数据库无响应时,自动分发到下一个候选连接,以实现 failover。

[0048] 作为优选,在步骤(1.3)中,为避免FENCE慢或者FENCE不成功,还需要关闭数据库服务器操作系统的acpi服务,同时开启idrac的ipmi功能,并给用户赋予ipmi可开关机的OPERATOR角色。

[0049] 作为优选,在步骤(1.4)中,还关闭network服务的自动启动节点node1和node2;并增加该自动启动节点node1和node2的network服务启动项到rc.local。

[0050] 作为优选,在步骤(1.7)中,在数据库主节点和数据库镜像节点配置流复制密码文件;在数据库主节点启动数据库,添加replication数据库角色;在数据库主节点启动虚拟IP;在数据库主节点配置数据库访问控制文件pg_hba.conf,为了在集群脚本中要用到更新sky_pg_cluster数据库的表,则需要允许主节点、镜像节点、虚拟IP和仲裁节点以及回环地址通过sky_pg_cluster用户访问sky_pg_cluster数据库。

[0051] 作为优选,在步骤(1.10)中,插入初始数据,创建测试函数,用于测试数据库是否正常,包括所有表空间的测试,使用update不同的表空间中的数据,并不能立刻反映表空间的问题;因为大多数数据在shared_buffer中,如果表空间对应的文件系统io有问题,那么在checkpoint时会产生58类的错误,使用pg_stat_file函数来立刻暴露io的问题。

[0052] 作为优选,在步骤(1.11)中,配置心跳需要的密码文件,集群failover软件的sky_pg_clusterd.sh将用这个密码文件分别用于访问虚拟IP上的PostgreSQL监听端口、本机standby的PostgreSQL监听端口和VOTE_HOST上的跳转端口。

[0053] 本发明能够达到如下效果:

[0054] 1、本发明与传统方法相比,本发明实现了数据库层面完全无单点故障的风险,并且在连接层面使用会话保持来解决应用感知的问题,实现对应用完全透明的高可用 failover。

[0055] 2、在本发明的方法中,当主库数据库本身或数据库所在的硬件,或者存储故障,或者存储空间不足等问题发生时,数据库心跳检测程序将检测到问题,使用fence设备截断主

库与外界的联系,然后激活主库镜像并切换IP。

[0056] 3、在本发明的方法中,连接池和应用程序建立的TCP连接不受failover影响,因此这个会话不会中断。连接池和数据库的连接断开后自动连接。业务层在切换期间的会话自动回滚。

[0057] 4、本发明的方法中,在连接池或驱动层配置好多个对应的后端数据库连接,根据配置的算法,数据插入的SQL自动分发到后端数据库以实现负载均衡。当后端数据库无响应时,自动分发到下一个候选连接,以实现failover。

附图说明

[0058] 图1是本发明读写混合场景的一种架构原理示意图。

[0059] 图2是本发明写场景的一种架构原理示意图。

[0060] 图3是本发明HA切换软件实现的一种逻辑流程原理示意图。

[0061] 图4是本发明读写分离的一种逻辑流程原理示意图。

[0062] 图5是本发明写负载均衡以及HA的一种逻辑流程原理示意图。

[0063] 图6是本现有技术使用存储硬件复制构建的数据库HA系统的一种架构原理示意图。

[0064] 图7是本现有技术使用共享存储构建的数据库HA系统的一种架构原理示意图。

具体实施方式

[0065] 下面通过实施例,并结合附图,对本发明的技术方案作进一步具体的说明。实例一:PostgreSQL高弹性的高可用及负载均衡实现方法,参见图1、图2所示,读写混合场景实现方法和写场景实现方法。

[0066] (1) 读写混合场景实现方法步骤如下:

[0067] (1.1) 准备3台服务器,其中2台服务器分别用于数据库主节点和数据库镜像节点,另1台服务器用于仲裁节点。

[0068] (1.2) 分别将数据库主节点服务器的主机、数据库镜像节点服务器的主机和仲裁节点服务器的主机配置在同一个网段,并设数据库主节点为node1,设数据库镜像节点为node2,设数据库的一个虚拟IP为虚拟IP1,设数据库的另一个虚拟IP为虚拟IP2,设仲裁节点为VOTE_HOST。

[0069] (1.3) 配置数据库主节点和数据库镜像节点的FENCE设备,为了防止发生脑裂现象,让FENCE设备在自动failover前关闭服务器。为避免FENCE慢或者FENCE不成功,还需要关闭数据库服务器操作系统的acpi服务,同时开启idrac的ipmi功能,并给用户赋予ipmi可开关机的OPERATOR角色。

[0070] (1.4) 配置数据库主节点和数据库镜像节点的网络,编写虚拟IP配置文件。

[0071] (1.5) 在数据库主节点和数据库镜像节点安装PostgreSQL数据库软件。

[0072] (1.6) 在数据库主节点初始化数据库。

[0073] (1.7) 在数据库主节点配置数据库访问控制配置文件pg_hba.conf、postgresql.conf和recovery.conf,这三个数据库访问控制配置文件都用于数据库复制、启动数据库、创建流复制用户和启动虚拟IP;在数据库主节点和数据库镜像节点配置流复

制密码文件。在数据库主节点启动数据库,添加replication数据库角色。在数据库主节点启动虚拟IP。在数据库主节点配置数据库访问控制文件pg_hba.conf,为了在集群脚本中要用到更新sky_pg_cluster数据库的表,则需要允许主节点、镜像节点、虚拟IP和仲裁节点以及回环地址通过sky_pg_cluster用户访问sky_pg_cluster数据库。

[0074] (1.8) 创建镜像数据库,配置数据库镜像节点流复制环境,并启动镜像虚拟IP。

[0075] (1.9) 配置仲裁节点,设置用于仲裁网络异常的数据监听端口,即配置仲裁机,在仲裁机上起一个监听端口,用于跳转到虚拟IP的PostgreSQL主库监听端口。

[0076] (1.10) 在数据库主节点创建心跳检查函数,数据库主节点的变更将自动复制到数据库镜像节点。插入初始数据,创建测试函数,用于测试数据库是否正常,包括所有表空间的测试,使用update不同的表空间中的数据,并不能立刻反映表空间的问题;因为大多数数据在shared_buffer中,如果表空间对应的文件系统io有问题,那么在checkpoint时会产生58类的错误,使用pg_stat_file函数来立刻暴露io的问题。

[0077] (1.11) 编写心跳检查软件和仲裁节点检查软件,并把心跳检查软件部署在数据库主节点和数据库镜像节点,把仲裁节点检查软件部署在仲裁节点,用于自动failover。配置心跳需要的密码文件,集群failover软件的sky_pg_clusterd.sh将用这个密码文件分别用于访问虚拟IP上的PostgreSQL监听端口、本机standby的PostgreSQL监听端口和VOTE_HOST上的跳转端口。

[0078] (1.12) 开启用于自动failover的心跳检查软件。

[0079] (1.13) 配置组件监控检查,所述组件包括用于负责心跳检测和故障时激活镜像以及切换虚拟IP的HA切换软件,所述组件还包括用于读写负载均衡和客户端会话保持连接池的中间件,所述组件还包括用于数据库主节点IP读写连接和数据库镜像节点IP读连接的虚拟IP,所述组件还包括用于投票解决单边网络故障的仲裁节点,所述组件还包括用于解决脑裂问题的FENCE设备,所述组件还包括用于负责读写的数据库主节点和从数据库主节点获取实时增量同步且只负责读的数据库镜像节点。

[0080] 所述HA切换软件的实现过程是:参见图3所示,

[0081] (1.13.1) 启动,判断主节点是否正常。

[0082] (1.13.2) 如果主节点不正常则结束,如果主节点正常则判断本地节点角色是否为镜像节点。

[0083] (1.13.3) 如果本地节点不是镜像节点则结束,如果本地节点为镜像节点则计数清零,并检查主节点状态。

[0084] (1.13.4) 如果主节点状态健康则结束,如果主节点状态不健康则检查仲裁节点状态。

[0085] (1.13.5) 如果仲裁节点状态不健康则再次进行计数清零,如果仲裁节点状态健康则从仲裁节点检查主节点状态。

[0086] (1.13.6) 如果从仲裁节点检查主节点状态为健康状态则再次进行计数清零,如果从仲裁节点检查主节点状态为不健康状态则进行主节点累计异常次数自增计数,然后进行主节点连接异常次数判断1。

[0087] (1.13.7) 如果主节点连接异常次数判断1是第一次时则检查镜像节点延迟是否正常,如果主节点连接异常次数判断1不是第一次时则进行主节点连接异常次数判断2。

[0088] (1.13.8) 如果检查镜像节点延迟为不正常则再次进行计数清零,如果检查镜像节点延迟为正常则生成镜像节点健康状态标记,然后则进行主节点连接异常次数判断2。

[0089] (1.13.9) 如果主节点连接异常次数判断2的异常次数未达到阈值则再次进行计数清零,如果主节点连接异常次数判断2的异常次数达到阈值则检查镜像节点状态标记。

[0090] (1.13.10) 如果检查镜像节点状态标记为正常则激活镜像、切换虚拟IP和生成切换标记,并在激活镜像、切换虚拟IP和生成切换标记后结束。如果检查镜像节点状态标记为不正常则生成主节点不健康标签和不发生切换,在生成主节点不健康标签和不发生切换后结束。

[0091] (1.14) 配置用于读写分离的连接池。参见图4所示,所述读写分离的实现过程是:APP提交SQL来判断SQL是读还是写,如果判断SQL是写则分发给主节点,如果判断SQL是读则分发给镜像节点。

[0092] (1.15) 人为的switchover测试,关闭数据库主节点,关闭虚拟IP1,激活数据库镜像节点数据库,数据库镜像节点数据库recovery.conf自动修改为recovery.done,切换为主角色,在数据库镜像节点启动虚拟IP1,数据库主节点recovery.done改为recover.conf,启动数据库主节点,角色切换为镜像角色。

[0093] (1.16) 通过拔数据库主节点网线,或者关闭数据库主节点,或者关闭数据库主节点服务器,或者关闭数据库主节点网卡来自动进行failover测试。

[0094] (1.17) 当failover发生时,数据库镜像节点自动切换为数据库主节点,同时虚拟IP1自动在数据库镜像节点开启。

[0095] (2) 写场景实现方法步骤如下:

[0096] (2.1) 准备至少2台用于安装数据库的主机。

[0097] (2.2) 在每台主机上安装数据库软件。

[0098] (2.3) 在每台主机执行初始化数据库操作,并配置数据库监听端口。

[0099] (2.4) 在每台主机创建同样的SCHEMA、用户、密码。

[0100] (2.5) 在每台主机创建同样的业务表和约束。

[0101] (2.6) 配置DNS,使主机名对应多个数据库主机的IP,同时配置监控,当数据库节点发生变更时实时更新DNS条目。

[0102] (2.7) 配置连接池或驱动来实现写负载均衡以及HA。参见图5所示,所述写负载均衡以及HA的实现过程是:APP提交SQL,由SQL分发策略采用写的方式分发给节点1,分发给节点1后判断分发是否成功,如果分发成功则分发完成,如果分发不成功则SQL分发策略尝试分发给下一个节点,直到分发成功为止则实现分发完成作业。

[0103] (2.8) 应用软件通过pgbouncer连接数据库,DNS策略配置为round-robin模式,当新建连接时,轮询的选择DNS解析出来的IP,从而POOLSERVER分布在不同的后端数据库,实现负载均衡。

[0104] 本实例对于读写混合场景是,利用PostgreSQL数据库实时流复制实现双份数据库镜像,解决依赖存储的高可用的问题。自定义的数据库心跳和切换逻辑实现数据库failover,解决购买商业高可用软件的问题。使用仲裁和fence设备解决脑分裂问题。使用连接池会话层保持解决应用感知问题。连接池可部署多个,不存在单点故障。同时连接池可实现读写分离。同步和异步完全由事务决定,程序如果发起同步事务,则事务必须到达镜像

节点后才提交完成；如果是异步事务，则事务到达主节点后就可以提交；解决了重要事务绝对不丢失，不重要事务选择异步则减少性能损失；实现了弹性控制。对镜像节点的延迟监控，如果发现镜像延迟达到阈值，告警。为防止数据丢失，在镜像延迟超过阈值时不会发生切换。本实例对于写场景是，记录用户行为数据。使用完全独立的两套硬件，创建相同的 SCHEMA，实现写负载均衡。在连接池或驱动层实现高可用，负载均衡，解决了连接层单点。当主库数据库本身或数据库所在的硬件，或者存储故障，或者存储空间不足等问题发生时，数据库心跳检测程序将检测到问题，使用fence设备截断主库与外界的联系，然后激活主库镜像并切换IP。连接池和应用程序建立的TCP连接不受failover影响，因此这个会话不会中断。连接池和数据库的连接断开后自动连接。业务层在切换期间的会话自动回滚。在连接池或驱动层配置好多个对应的后端数据库连接，根据配置的算法，数据插入的SQL自动分发到后端数据库以实现负载均衡。当后端数据库无响应时，自动分发到下一个候选连接，以实现 failover。

[0105] 上面结合附图描述了本发明的实施方式，但实现时不受上述实施例限制，本领域普通技术人员可以在所附权利要求的范围内做出各种变化或修改。

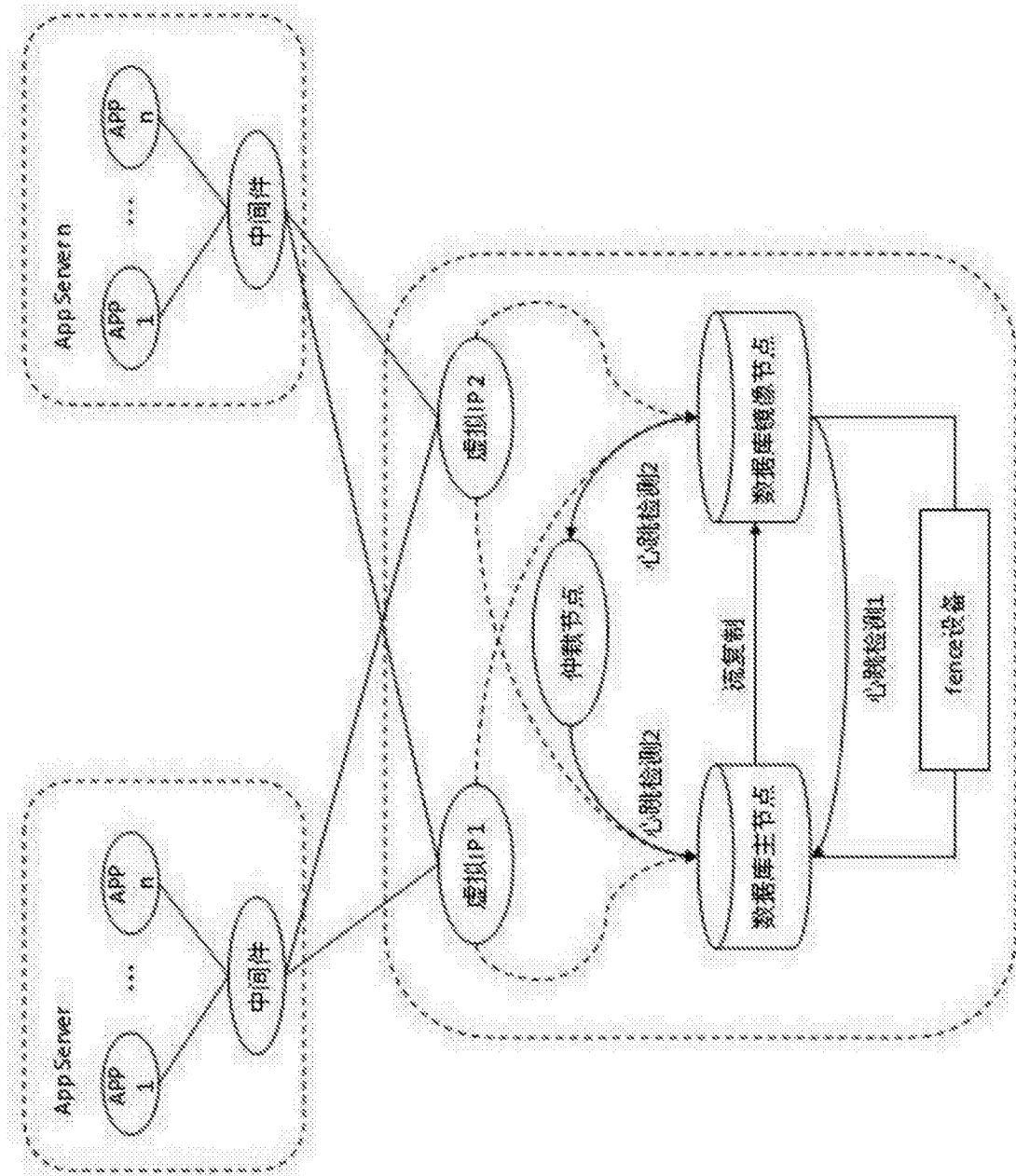


图1

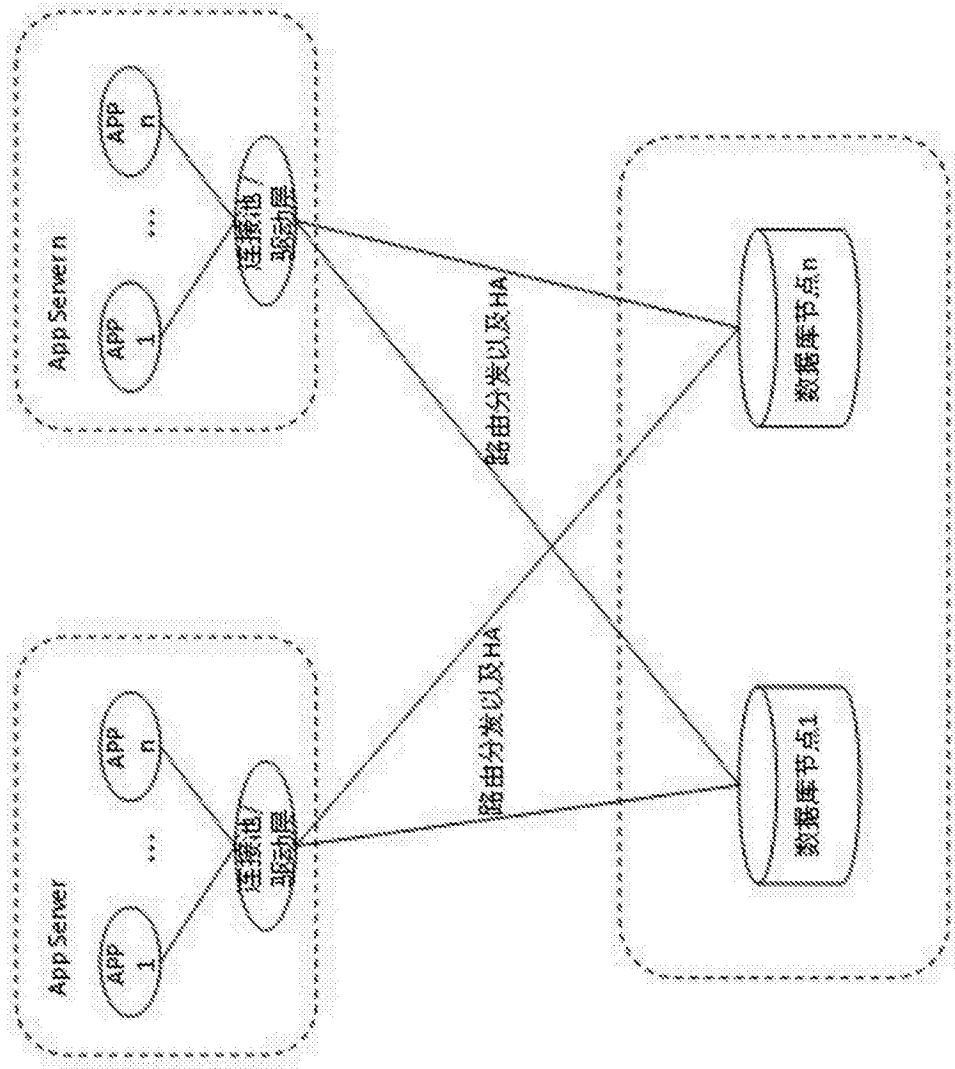


图2

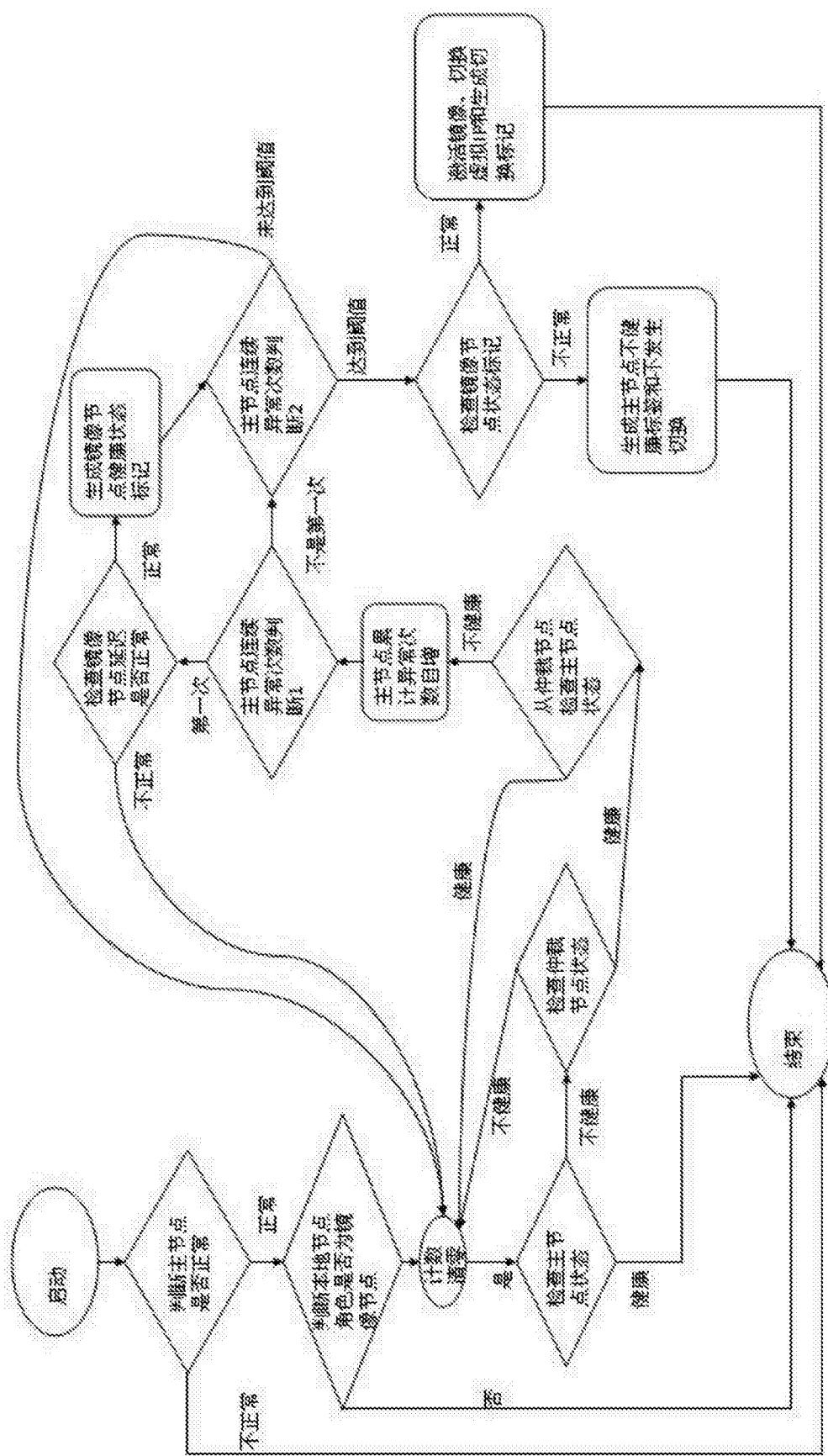


图3

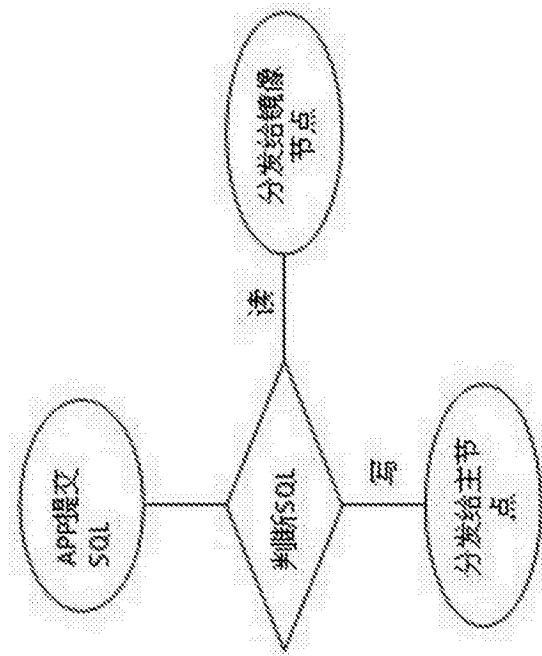


图4

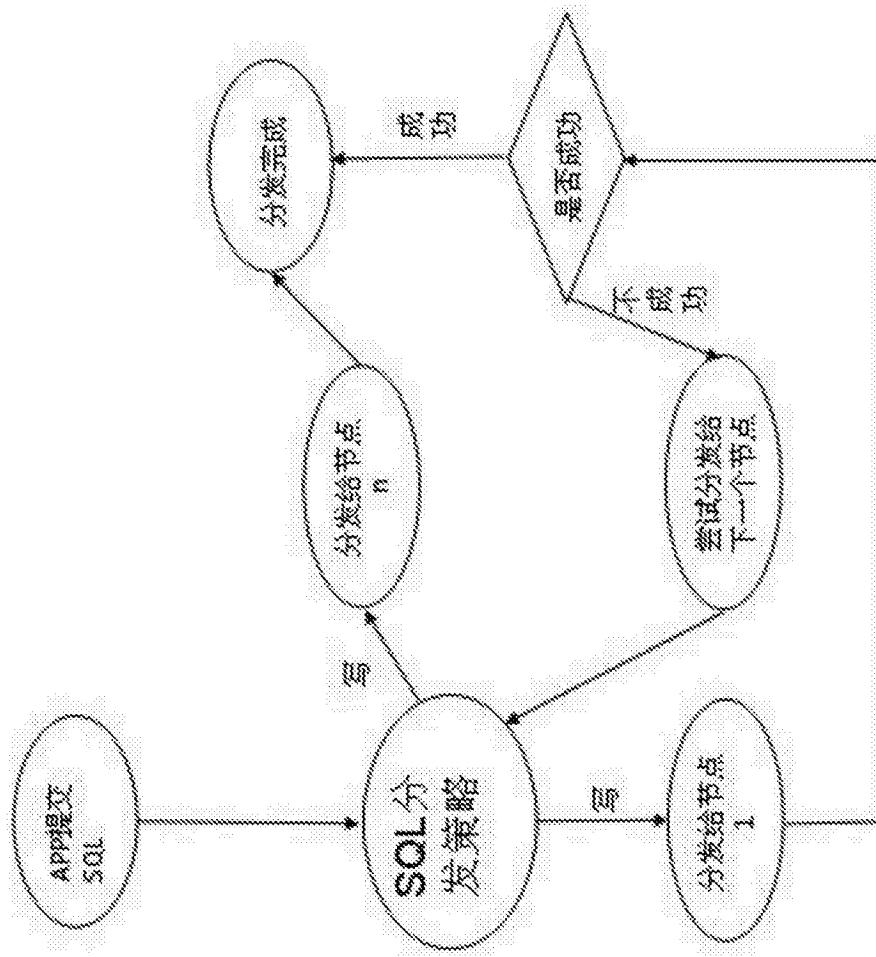


图5

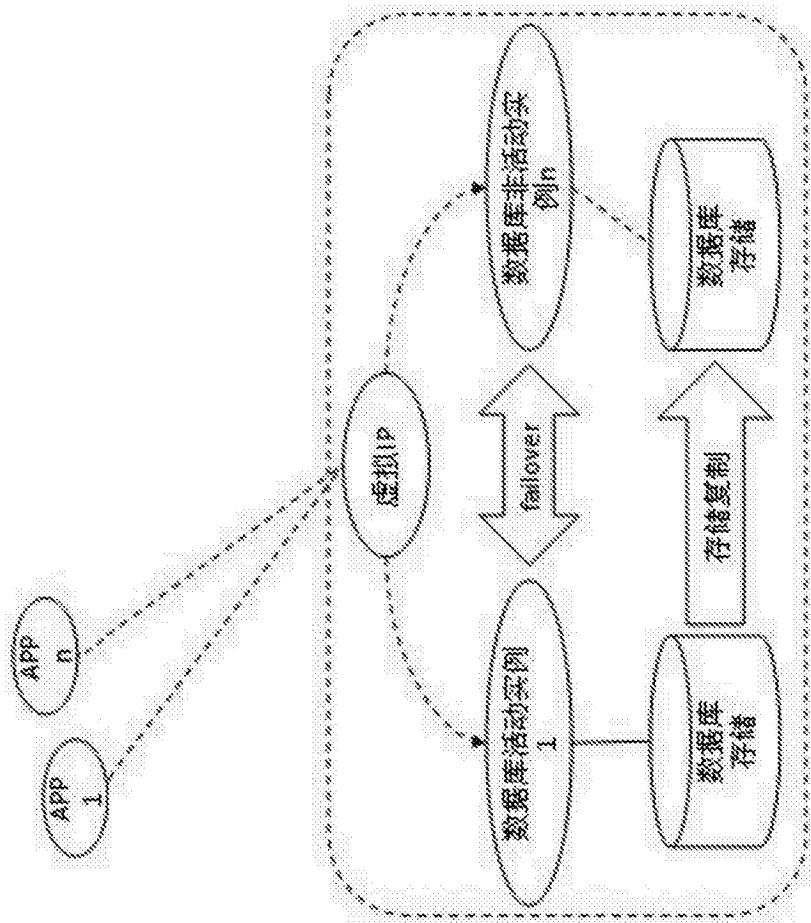


图6

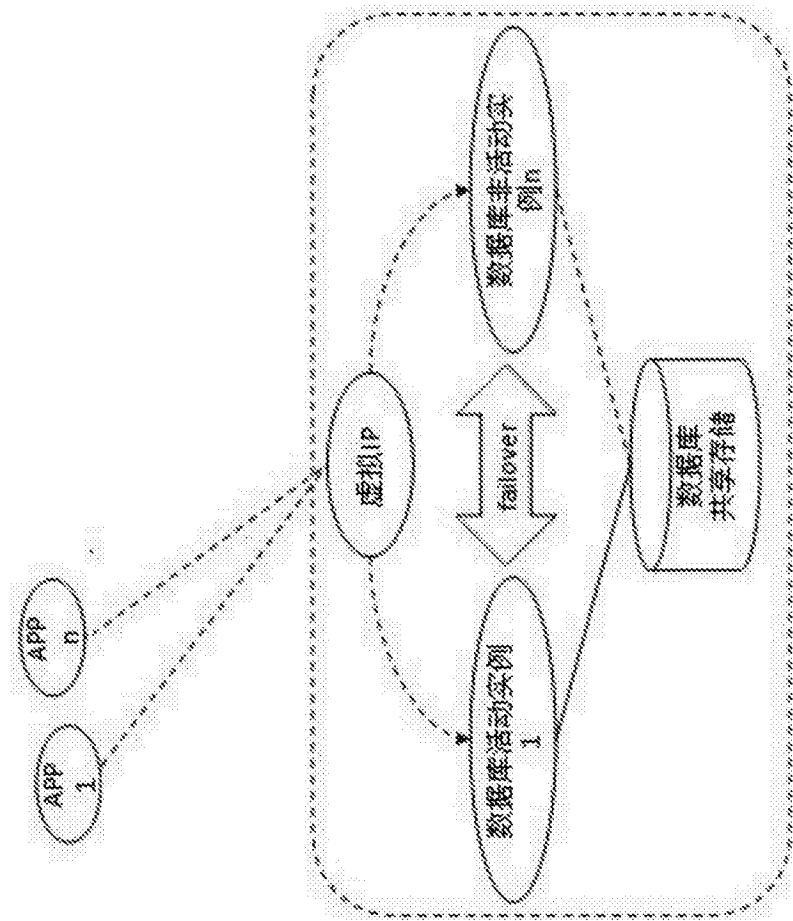


图7