DAOWOO
达 沃 时 代

PDB数据同步与数据恢复功能测试、PDB性能测试总结

2016.09

涂仲秋 田末

目录

PDB数据同步与数据恢复功能介绍

- 数据同步的目的
  - 多节点间保持数据一致，相互数据备份
  - PDB功能高可用
  - 提高访问效率

- 数据同步的方法
  - 稳态时（3节点正常工作）：严格两阶段提交，先Prepare阶段，后Commit阶段
  - 非稳态时（发生节点离线、重启后重新加入集群）：
    - 3种同步模式：
      - DIFF
      - LOGS
      - SNAP

测试点

- 正常场景
  - 3节点monitor集群正常工作情况下，下发一个或多个写请求命令，命令响应成功后检查各节点上的KV和SQL数据内容是否保持一致；

- 异常场景
  - 3节点monitor集群正常工作情况下，构造节点离线、重启的场景，检查异常发生后各阶段各节点上的KV和SQL数据库内容是否保持一致；
    - 比如，构造一个Follower节点离线的场景，该节点离线后，下发多次写请求命令修改数据库内容，检查在该节点离线期间剩余两节点的数据同步是否正常；当该节点重新上线后，检查该节点的数据同步是否正常；
    - 构造Leader节点和一个Follower节点重启的场景，这两个节点重启后，检查PDB服务是否已关闭，当两节点完成重启重新上线后，检查3节点上的数据是否保持一致；

测试困难与解决方法

DAOWOO 达沃时代

- 如何判断各节点上的数据保持一致？
  - KV是内存数据库，内容不可见；
  - SQLite 内容虽然可见，但数据量大了之后一一比对很困难；

  方法：通过比较各节点上的事务日志，对于写操作，各节点在更新数据库之前会都先对此事务记录事务日志，如果各节点上事务日志都保持一致，可以认为各节点上的数据内容也保持一致（各事务都已提交到数据库）。

- 如何比较各节点上的事务日志？
  - 各节点上的事务日志分两种，KV数据库事务日志和SQL数据库事务日志，分别记录；
  - 事务日志是一个二进制文件，由要记录的内容序列化后生成，不可读且无法比较；

  方法：做一个事务日志解析工具（兼容KV、SQL），把原始事务日志文件里记录的内容解析出来，并把解析出的内容写入到一个新的文本文件里，便于阅读与比较。

- 如果使用快照同步数据，如何判断这个同步过程是否正确？
  - 使用快照时，接收快照的节点不会记录事务日志，无法通过事务日志判断数据同步是否成功；
  - 快照文件也是一个二进制文件，序列化后内容无法识别

  方法：做一个快照解析工具，将生成的快照文件内容解析出来，并把解析出的内容写入到一个新的文本文件中，便于比较与阅读；

- 如何下发写操作命令改变数据库内容？
  - 为保证测试环境可控，测试环境只安装了3节点的monitor及CLIServer，其他服务都没有安装

  方法：使用 yfslog set_quota --capacity 命令设置日志配额，这条命令下发后，会涉及到KV和SQL数据库中内容的修改，且在测试环境下没有其他模块可以下发类似操作，可以作为测试中的写请求命令下发；

测试实施及典型问题定位

DAOWOO 达沃时代

- 问题一：3节点正常工作情况下，下发写请求后解析各节点事务日志时发现，Leader节点上事务日志CRC校验和前后计算不一致。
  - 这里不一致指的是：PDB在记录事务日志时计算出的CRC校验和与解析工具在解析日志时计算出的校验和不一致；且PDB计算出的校验和均为 0xFFFFFFFF；
  - 定位过程：
    - 1、gdb单步跟踪计算CRC校验和的过程，发现计算函数 mycrc32 在计算校验和过程中使用了大量宏运算，单步跟踪时看不出问题所在；
    - 2、阅读CRC校验和计算模块代码，发现有一个校验和计算初始化函数 mycrc32_init，用于初始化CRC计算表，而PDB模块在计算CRC校验和时没有使用这个初始化函数，找到问题线索；
    - 3、检查其他使用这个CRC计算公共模块的代码，发现都在使用 mycrc32 计算校验和之前，使用 mycrc32_init 做了计算表的初始化，进一步确认问题原因；
    - 4、修改PDB代码，在模块初始化节点使用 mycrc32 初始化CRC计算表，然后再检查各日志校验和，不再出现计算不一致以及计算出的校验和均为 0xFFFFFFFF的情况，问题确认；

- 问题二：构造Leader节点离线场景，离线期间下发多次写请求命令，该节点重新上线后，发现该节点的KV和SQL事务日志都没有同步成功。
  - 定位过程：
    - 1、怀疑可能同步过程还未完成，反复检查事务日志文件，发现仍然没能同步；
    - 2、检查该节点和新选举出Leader节点上的monitor日志，发现新Leader已经启动并完成了两个数据库的同步，该节点上的monitor日志也显示该节点收到了Leader发生的同步信息，除了事务日志没有同步，其他信息都显示正常，看不出任何问题；
    - 3、经过与路俊讨论，最终确定了问题所在：PDB在记录事务日志时，使用的是标准库中basic_ostream的子类 basic_Logofstream 将事务日志内容写入磁盘，写入时会先判断写入数据的长度，如果长度不够，会先把要写入磁盘的数据保存在 basic_Logofstream 的缓冲区中，这里我们检查事务日志时发现事务日志没有同步就是因为待同步的内容还保存在basic_Logofstream 的缓冲区中，没有刷入磁盘。
    - 4、问题基本定位后，修改代码，当该节点（Follower）收到Leader发送过来的日志同步信息后，为保证所有同步信息不再丢失，Follower第一时间把同步的日志内容刷入磁盘。

- 问题三：构造Leader和一个Follower节点重启场景，重启完成后发现有一个Follower的SQL日志没有同步完成（比Leader上的日志少），再次下发写命令才能同步完成（与Leader一致）。
  - 定位过程：
    - 1、从问题现象来看，这个问题与上一个日志没有同步的问题很类似，现象基本一致；
    - 2、检查monitor日志，发现日志未同步的Follower是采用LOGS模式与Leader同步日志，这与上一个问题不一样，上一个问题是采用 DIFF 模式与Leader同步日志，找到问题线索；
    - 3、跟踪代码中 LOGS 模式下Follower同步数据时的处理函数，发现同样没有把收到的日志信息强制刷盘，问题基本定位；
    - 4、修改代码，LOGS模式下Follower收到Leader发送的同步日志信息时，采取强制刷盘，编译版本后重新验证此问题出现的场景，SQL日志不同步的问题得到解决；

测试结果

- 共执行了29个测试用例，覆盖了不同数目节点离线、重启、离线加重启场景下，各个阶段集群中各节点数据同步和数据恢复功能的测试；
- 测试中总共发现了8个场景下数据同步与数据恢复过程出现异常，经过定位及修复后验证，所发现的问题已经全部解决；

小结

- 通过编写的几个工具：事务日志解析工具、快照解析工具、扩展yfssys queryleader命令功能、编写断网脚本等显著提高了测试的效率，帮助我们快速发现问题及解决问题；

PDB基础性能测试介绍

- KV数据库性能测试方案
  - 插入不同数量级KV Entry所需时间
  - 在不同数量级KV Entry下插入不同数量级KV Entry所需时间
  - 在不同数量级KV Entry下读取不同数量级KV Entry所需时间
  - 在不同数量级KV Entry下删除不同数量级KV Entry所需时间
  - 在不同数量级KV Entry下修改不同数量级KV Entry所需时间

- SQLite数据库性能测试方案
  - 存储介质
    - SSD
    - SATA
  - 测试用例设计
    - insert方式插入不同数量级记录时间
    - replace方式插入不同数量级记录时间
    - 在不同数量级记录基础上insert方式插入不同数量级记录时间
    - 在不同数量级记录基础上replace方式插入不同数量级记录时间
    - 在不同数量级记录基础上读取不同数量级记录所需时间
    - 在不同数量级记录基础上删除不同数量级记录所需时间
    - 在不同数量级记录基础上更新不同数量级记录所需时间

- 测试方法
  - 把测试代码单独编译成可执行文件，接收3~5个命令行参数，可覆盖所有测试用例
  - 编写一个shell脚本自动化执行测试用例，每个用例执行5次，脚本读取每次测试后输出的日志文件，读取测试结果（测试时间、数据库文件大小），然后在5个测试结果中去掉最大值、最小值后剩余3个值取平均作为最终结果

- SQLite性能优化
  - 精简SQL语句执行过程，尽可能少的调用API sqlite3_exec()，原先的做法调用了4次。
  - 优化效果：
    - insert 操作在SATA盘上依数量级不同性能可提升 11% ~ 27%
    - insert 操作在SSD盘上依数量级不同性能可提升 3% ~ 9%

Thanks !

联系我们

北京总部
北京市朝阳区安定路一号国家奥林匹克体育中心体育场东南门2层2160室
邮编：100029
电话：+8610 8843 7583
传真：+8610 8437 6490

武汉分公司
武汉市洪山区珞瑜路33号中部创意大厦1907、1908
邮编：430074
电话：+8627 8786 2881
传真：+8627 8786 2567