新浪微博开放平台 Redis 实践

@唐福林
http://weibo.com/tangfl
http://fulin.org

大纲

- Redis 简介
- 新浪微博中的Redis实践
  - 好友关系
  - 计数器
- 经验教训

Redis

- in memory (database?)
- data can dump to disk
- many useful data structure
- FAST both read and write
- we start use from 2.0, now 2.4

微博=feed+关系+数字

微博=feed+关系+数字

- feed
  - 看微博，看评论
  - 发微博，转发微博，发评论

微博=feed+关系+数字


唐福林
97 关注
1654 粉丝
1746 微博
首页
提到我的
评论
私信
收藏
相册
音乐
活动
投票
微博精选
风云榜
更多»
有什么新鲜事想告诉大家？
老人跌倒干预指南
表情
图片
视频
音乐
话题
投票
发布
家庭
我的微群
猜你喜欢
全部
原创
图片
视频
音乐
查找作者、内容或标签
高级搜索
leeyanva：哈哈，都是微薄惹的祸@唐福林
@全球热门伤不起：世界上最远的距离，不是生与死的距离。而是我站在你身边，你他娘的却在玩手机。
9月6日 16:06 来自新浪微博
转发(1377) | 评论(183)
9月6日16:32 来自新浪微博
转发 | 收藏 | 评论
新版微博小贴士 体验问卷
可能感兴趣的人
换一换
招财艳艳
+加关注
我们有8共同好友
我的好友中：阿里大哥、朱倩、赵鹏城等8 人也与她互相关注
风是云的爱人
+加关注
我们有8共同好友
邓小征
+加关注
我们有11共同好友
推荐/隐私设置
更多»
热门话题
换一换
又是一年迎新时 (12876682)
老人跌倒干预指南 (36467)
老人跌倒是否扶起有标准了，你能hold住么

微博=feed+关系+数字

- 关系
  - 关注，取消关注
  - 关注列表，粉丝列表

微博=feed+关系+数字


关注/粉丝
关注
全部关注(97)
互相关注(55)
粉丝(1654)
邀请站外好友
找人
猜你喜欢
我关注了97人
输入昵称或备注
全部
互相关注
未分组
家庭
技术
同事
更多
创建分组
排序方式:
关注时间
最近更新
昵称首字母
最近联系
粉丝数
艳博
未分组
离线
北京,海淀区
粉丝7510人
//@吕宏伟: 挖槽，你们可别让他们堵。 上次堵一下，限号了。又堵一下，抽号了。 再堵一下不得蹲号? //@漂泊云间：//@小户行云海肴:恩。 // @任志强 :特权的伤害。 // @王巍w :教主这经典几句必须转起来，杠杠的！ // @赵丽华 :昨天从长虹桥到甘家口，活生生堵了三个小时。。 // @湖南十年： // @岁荣(今天 18:00)
开源中国 (设置备注)
技术
取消关注
离线
广东,广州
粉丝5744人
tomcat-redis-session-manager 是一个用来将 Tomcat 的 Session 数据存储在 Redis 库中的项目。使用方法请看 http://t.cn/a1W5YV(今天 15:17)
zolker(杨尚刚)
同事
在线
北京,海淀区
粉丝154人
说得好[给力](9分钟前)
阿里大哥(建新)
同事
在线
北京,
粉丝3165人
眼神很好！(今天 19:55)
可能感兴趣的人
换一换
老黄
加关注
我们有8共同好友
我的好友中：李嵩波、sunli1223、TreapDB等8人也与他互相关注
张兰兰
加关注
我们有10共同好友
呛呛cei-
加关注
我们有12共同好友
更多>>
管理我的黑名单

微博=feed+关系+数字

- 数字
  - 微博，粉丝，关注数
  - 评论给我的，@我的，我评论的
  - 小黄签提醒：新粉丝，新@，新评论
  - 未读微博数

微博=feed+关系+数字

微博=feed+关系+数字

- 数字
  - 微博，粉丝，关注数
  - 评论给我的，@我的，我评论的
  - 小黄签提醒：新粉丝，新@，新评论
  - 未读微博数

微博=feed+关系+数字


共196条
leeyanva：哈哈，都是微薄惹的祸@唐福林
@全球热门伤不起：世界上最远的距离，不是生与死的距离。而是我站在你身边，你他娘的却在玩手机。
9月6日 16:06 来自新浪微博
转发(1376) | 评论(182)
9月6日16:32 来自新浪微博 | 举报
转发 | 收藏 | 评论
首页
提到我的
@我的微博
@我的评论
评论
私信
收藏
@使用小帮助
Q1：什么是@提醒?

微博=feed+关系+数字

- 数字
  - 微博，粉丝，关注数
  - 评论给我的，@我的，我评论的
  - 小黄签提醒：新粉丝，新@，新评论
  - 未读微博数

微博=feed+关系+数字


微博小黄签 设置哪些新消息，通过微博小黄签提醒我
新评论提醒
设置哪些评论计入评论提醒数字中
评论的作者是: 所有人 关注的人
新增粉丝提醒
新私信提醒
@提到我提醒
设置哪些@提到我的微博/评论计入@提醒数字中
微博/评论的作者是: 所有人 关注的人
微博的类型是: 所有微博 原创的微博
群内新消息提醒
相册新消息提醒
新通知提醒
新邀请提醒
位置示意:
我的微博
消息
帐号
12位新粉丝，查看粉丝
31234条新评论，查看评论
230条微博/评论@我，查看@
156条群内新消息，查看消息
查看私信
查看通知

微博=feed+关系+数字

- 数字
  - 微博，粉丝，关注数
  - 评论给我的，@我的，我评论的
  - 小黄签提醒：新粉丝，新@，新评论
  - 未读微博数

微博=feed+关系+数字


首页
提到我的
评论
私信
收藏
最新微博
我的微群
猜你喜欢
全部
原创
图片
视频
音乐
高级搜索
有 3 条新微博，点击查看
朱倩：我参与了投票【公司发月饼，该不该缴纳"月饼税"？】 http://t.cn/au2lGh 投给"不该" 这1个选项
我的好友中： 阿里大哥、朱倩、王兆源等11 人也与她互相关注
风是云的爱人
+ 加关注
我们有11共同好友
鹤舞西风

微博=feed+关系+数字

- feed
- mysql
- mc

微博=feed+关系+数字

- 关系

微博=feed+关系+数字

- mysql: relation.following
  - fromuid, touid, addtime
- 关注列表： select * from following where fromuid=? order by addtime desc
- 粉丝列表： select * from following where touid=? order by addtime desc
- 问题： fromuid, touid 都为索引，插入慢

微博=feed+关系+数字

- mysql: relation.following relation.follower
  - fromuid, touid, addtime
- 关注列表： select * from following where fromuid=? order by addtime desc
- 粉丝列表： select * from follower where touid=? order by addtime desc
- 问题：插入两张表，非事务，一致性

微博=feed+关系+数字

- 双向关系：关注与粉丝的交集
  - 实时计算：读的时候计算。
    - 问题：效率
- 预先计算：写的时候计算，存储
  - 问题：一致性，空间占用

微博=feed+关系+数字

- 我和ta的共同关注
- 我关注的人里有多少关注了ta
- 我的粉丝里有多少关注了ta
- 。。。

微博=feed+关系+数字

- 我们想要的
- 简单：c/java ， 可快速通读代码
- 可靠：经过验证的
- 高效：读写速度满足需要
- 方便实现需求

微博=feed+关系+数字

- redis
- hash :
  - key : user id
  - fields : friends ids
  - value : add time

微博=feed+关系+数字

- redis
- hash :
  - hset fromuid.following touid addtime
  - hset touid.follower fromuid addtime
  - hgetAll fromuid. following
  - hgetAll touid.follower ?
    - 姚晨粉丝 11,704,598 @Wed Sep 7 21:46:33 CST 2011

微博=feed+关系+数字

- hash-max-zip-size
  - 64 -> 256，节省近 1/3 内存
  - cpu 消耗增大
- hgetAll cost too much cpu
  - add mc
- high delay

微博=feed+关系+数字

唐福林

原来 redis 每隔一段时间出一堆耗时 50+ms 的请求，是 server 在 sync aof 文件哪。原版的 redis 没有主动控制 sync 的时间间隔，而是交给底层 os 去做，所以慢请求分布表现的比较随机

8月30日 17:39 来自新浪微博　　转发(1) | 删除 | 收藏 | 评论(9)

微博=feed+关系+数字

- redis
  - cache ? waste too much mem
  - storage ?
    - rdb may lost data
    - aof r/w too slow, recover too slow
    - all data in mem, waste money
- HA : master slave ? NO WAY
- memory fragment

微博=feed+关系+数字


唐福林

38G，已经开始 swap 了。48G 内存的机器上部署一个 redis 端口，极限是 38G //@唐福林: 38G 了，居然还没有崩溃，太难以置信了?

@唐福林： redis 的内存占用真是一个大坑哪，48g的机器，放4个端口，每个端口只能到8g；放2个端口，每个端口只能到18g；放一个端口，到34g了，可能马上就要出事故了吧

8月13日 08:27 来自iPhone客户端 转发(15) | 评论(11)

8月26日 11:09 来自新浪微博 转发(1) | 删除 | 收藏 | 评论(4)

微博=feed+关系+数字

- 更新
  - mysql binlog >> queue >> Java Processor >> redis
  - mysql binlog >> trigger >> redis

微博=feed+关系+数字

- 现状
- redis@weibo for now:
- TB 级
- growing fast

微博=feed+关系+数字

- 未来
- @摇摆巴赫 mysql modified + innodb
- following list of a user in one column
- still under dev

微博=feed+关系+数字

- 数字

微博=feed+关系+数字

- 永久计数
  - 用户的
    - 微博，粉丝，关注
    - @我的，我评论的，评论我的
  - 微博
    - 转发，评论

微博=feed+关系+数字


关注/粉丝
关注
全部关注(97)
互相关注(55)
粉丝(1654)
邀请站外好友
找人
猜你喜欢
我关注了97人
输入昵称或备注
全部
互相关注
未分组
家庭
技术
同事
更多
创建分组
排序方式:
关注时间
最近更新
昵称首字母
最近联系
粉丝数
艳博
离线
北京,海淀区
粉丝7510人
未分组
共196条
leeyanva：哈哈，都是微薄惹的祸@唐福林
@全球热门伤不起：世界上最远的距离，不是生与死的距离。而是我站在你身边，你他娘的却在玩手机。
9月6日 16:06
9月6日16:32
唐福林
原来 redis 每隔一段时间出一堆耗时 50+ms 的请求，是 server 在 sync aof 文件哪。原版的 redis 没有主动控制 sync 的时间间隔，而是交给底层 os 去做，所以慢请求分布表现的比较随机
8月30日 17:39 来自新浪微博
转发(1) | 删除 | 收藏 | 评论(9)
阿里大哥(建新)
在线
北京,
粉丝3165人
同事
眼神很好！(今天 19:55)
可能感兴趣的人
换一换
老黄
加关注
我们有8共同好友
我的好友中：李嵩波、sunli1223、TreapDB等8人也与他互相关注
首页
提到我的
@我的微博
@我的评论

微博=feed+关系+数字

- 临时计数
  - 小黄签提醒
    - 新粉丝，新@，新评论
- 未读微博
  - 聚合计算
  - 页面js每隔一段时间请求一次

微博=feed+关系+数字

- mc + queue + mysql
- 写入量：mysql 批量插入
- mc与mysql不一致
- 为了提高命中率，mc 空间需要足够大

微博=feed+关系+数字

- 我们想要的：
  - 大写入量
  - 大读取量
  - 持久化
  - 简单可靠高效

微博=feed+关系+数字

- redis
- k-v , 100 byte per k-v
  - mc 也一样
  - 单个业务十亿量级的数字个数
- hash , hget pipeline slow

微博=feed+关系+数字

- redis
- rdb ? may lost data
- aof ? grow too fast (4G/day)
- bgsave/bgrewriteaof influence parent

微博=feed+关系+数字

唐福林

redis bgsave 在 fork 出子进程的那一瞬间，以及后续子进程写磁盘的一段时间内，父进程对外提供的服务质量似乎都会受到影响。子进程写磁盘没有任何限速，导致磁盘带宽跑满，父进程又配置了 aof ，受影响是能理解的。但 fork 的那一瞬间，为什么也受影响呢？这个时候子进程还没有开始写文件呢

8月30日 11:43 来自新浪微博　　转发(2) | 删除 | 收藏 | 评论(6)

微博=feed+关系+数字

- redis rolling
  - 场景：微博的评论数
  - 每天写入亿级，每天读取十亿级
  - 明显的时间长尾
  - 目标：将一段时间前的 key 淘汰出去

微博=feed+关系+数字

- 现状
  - rediscounter @果爸果爸
    - array , not linked list
    - malloc all mem when start
    - hash key to position
    - write disk: asyn & slow down
    - add position to aof file

微博=feed+关系+数字

- 未来
- rediscounter + innodb
- auto roll cold data to disk

微博=feed+关系+数字

- 未读微博数
- @TimYang 亲自设计算法
- @XiaoJunHong 实现
- 向量相减
- redis (delay) - mc (throughput) - java hash map

经验教训

- 准确定位
- cache
- storage

经验教训

- 适用场景
  - 大量写入
  - 复杂数据结构
  - 简单数据结构+持久化
  - 容量小于内存

经验教训

- 容量规划
  - 容量增长预估
  - 读/写量预估
  - 数据结构
  - 内存碎片

经验教训

- 持久化
- 是否需要
- rdb or aof ?

经验教训

- 高延迟
- 持久化
- rehash

经验教训

- HA / Cluster
- 当前 Redis 本身支持不完美
- Jedis 客户端支持不完美

经验教训

- CPU 瓶颈
  - Redis 单线程
    - hset with big hash-max-zip-size
    - hgetAll
    - 对策： mc

经验教训

- 扬长避短
  - 内存操作快
  - 磁盘操作慢
  - 偶尔高延迟
  - 可能费内存

Thanks

@唐福林
http://weibo.com/tangfl
http://fulin.org

Q & A

PS. We are hiring !
contact me via @唐福林