ODF 2017 开源数据库论坛(北京)
OPEN-SOURCE DATABASE FORUM(BEIJING)

开源数据库正在改变世界

2017年8月24日-25日　北京－京仪大酒店

揭秘支撑美团点评万亿级访问的三大存储体系

赵应钢 美团点评高级技术专家

个人简介

- 美团点评服务运维部，高级技术专家
- 负责MySQL、Cellar、Squirrel的技术保障和服务支撑
- 曾任职于百度、新浪、去哪儿网，近10年专职DBA经验

目录

CONTENTS

三大存储发展现状

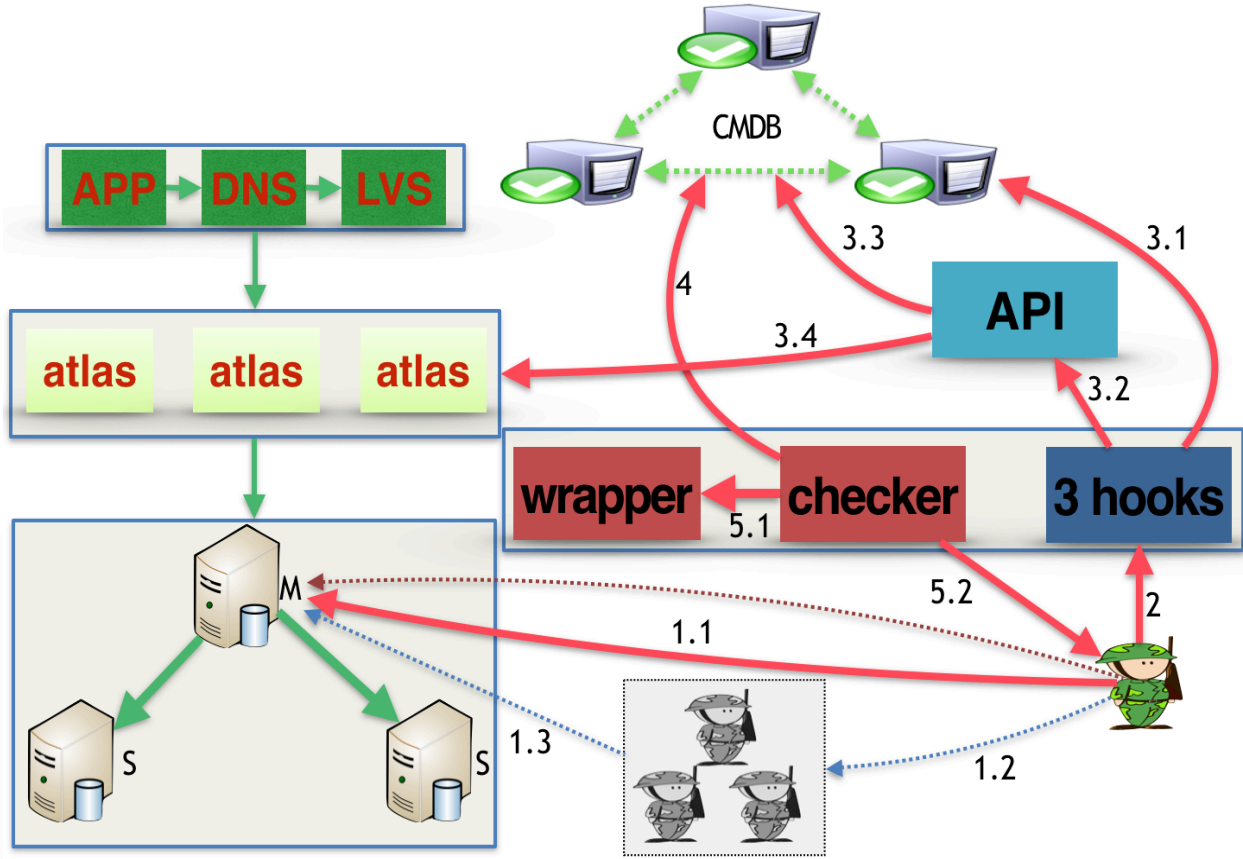
之MySQL： 存储的基石，老当益壮

MySQL服务现状

1. 消灭5.5，主力5.6，灰度5.7
2. 全面开启GTID+ROW+RC
3. 实例数以千计
4. 加速从RAID SSD过渡到Pcie

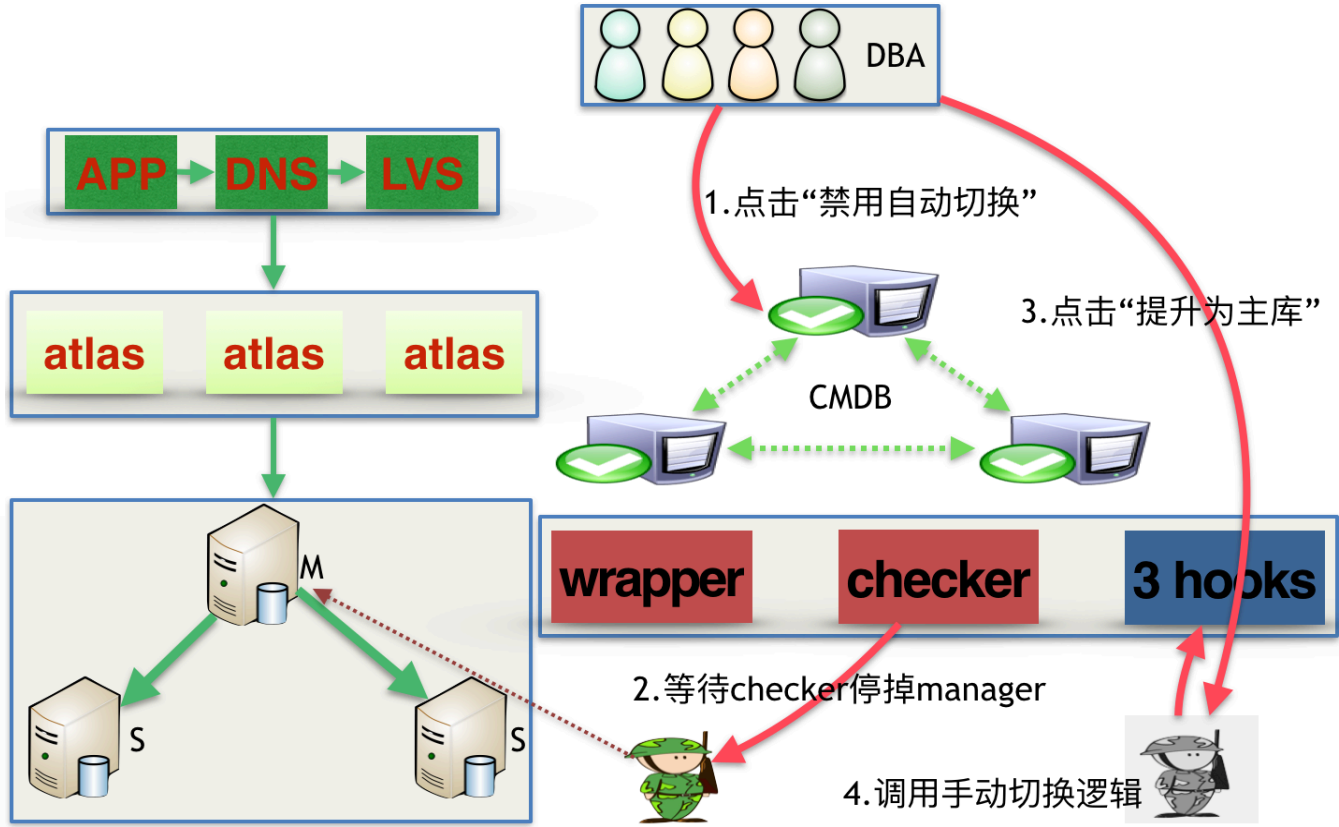
基于MHA改造的高可用系统-自动切换

1. 为社区贡献patch
2. 改造为按集群打印日志
3. 自动根据拓扑生成配置文件
4. 集群结构变动自动重写配置
5. 命令行和平台双保险
6. 异步后台进程防脑裂
7. GTID+semi-sync加强一致性
8. CMDB支持跨机房高可用



基于MHA改造的高可用系统-手动切换

1. 支持手动强制切换
2. 可配置权重
3. 同机房从库优先
4. OLTP从库优先
5. 检测和修复errant transactions



基于MMHA+proxy的流量管理

服务组【mtmop】创建日期【2016-03-01】 ⊕添加服务

基本信息

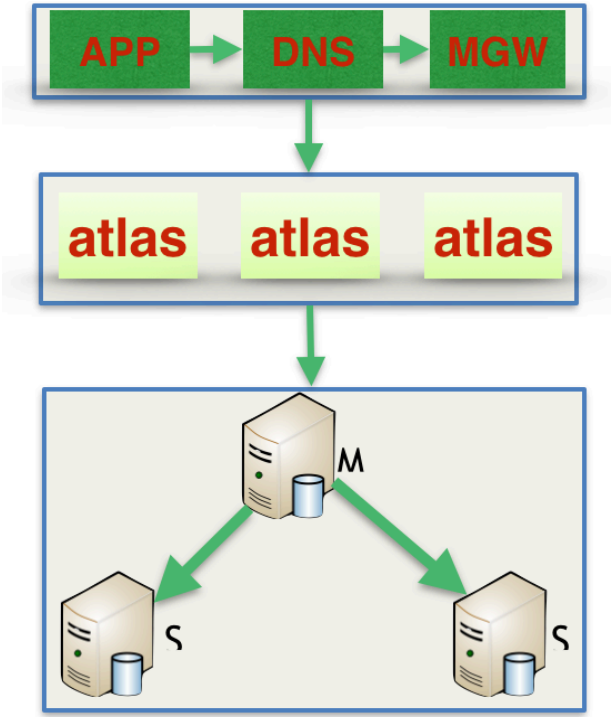
联系人	负责人	联系方式	描述信息	服务组类型	DBA联系人	DNS	是否涉及到钱	是否上线Atlas	是否上线Zebra	拓扑结构
[redacted]	[redacted]	[redacted]	mop	实时	[redacted]	查看	不涉及	需要上线	需要上线	查看

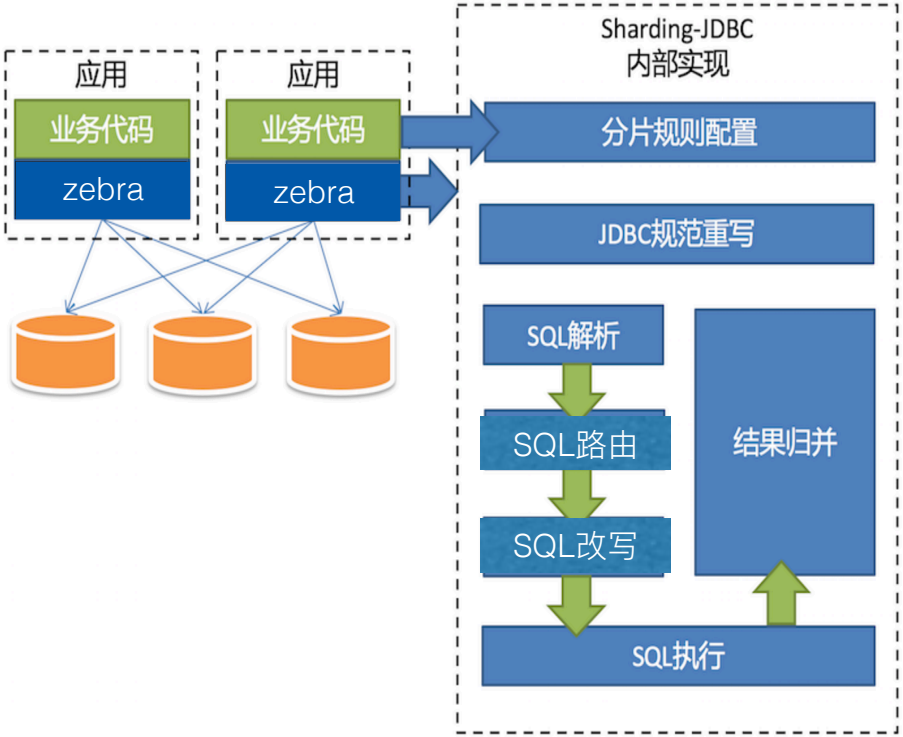
服务信息

mysql MHA切换状态: ◉ 启用 ○ 禁用 | 是否上半同步: ◉ 不启用 ○ 启用 | 是否上双1: ◉ 不启用 ○ 启用 Falcon-Screen | 查看服务组实时状态

实例	角色	配置	atlas参数	mha参数	实例参数	atlas操作	mha操作	实例操作
[redacted]mysql-[redacted]	master	虚拟机 内存: 32 GB 网卡: 双万兆 16核 磁盘: 0.52TB(SSD | Raid)	状态: online 角色: write,read	候选主库: yes 切换权重: 90	监控类型: mysql_cos	下线读 下线写	改参数	改监控 删除 编辑
[redacted]mysql-[redacted]	slave	虚拟机 内存: 32 GB 网卡: 双万兆 16核 磁盘: 0.52TB(SSD | Raid)	状态: online 角色: read	候选主库: yes 切换权重: 90	监控类型: mysql_cos	下线 标记维护	提升主库 改参数	改监控 删除 编辑
[redacted]mysql-[redacted]	slave	虚拟机 内存: 32 GB 网卡: 千兆 8核 磁盘: 0.52TB(SSD | Raid)	状态: online 角色: read	候选主库: yes 切换权重: 90	监控类型: mysql_cos	下线 标记维护	提升主库 改参数	改监控 删除 编辑

Proxy and Smart client





三大存储发展现状

之Squirrel：　　缓存为王，唯快不破

Squirrel服务现状

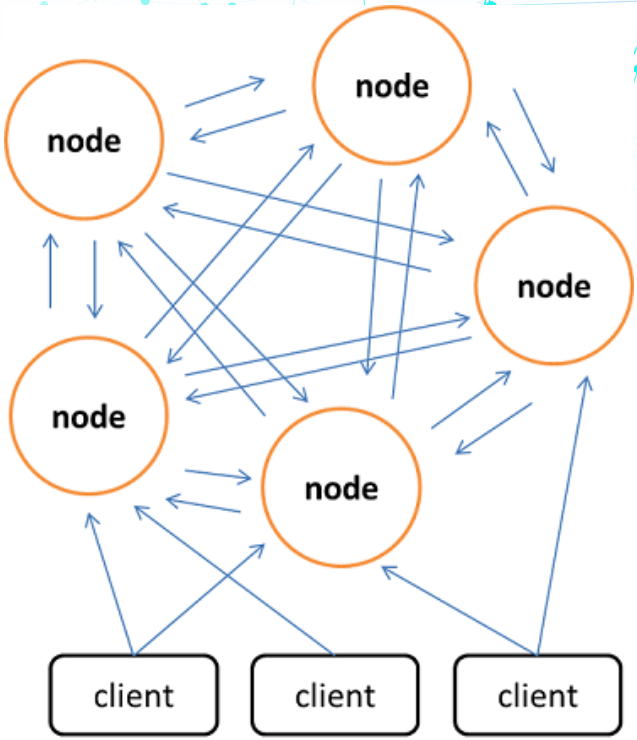
1. 服务端基于redis-cluster 3.0
2. 客户端基于jedis改造，与zk结合
3. 存储使用SAS物理机+docker
4. 调度基于hulk，实现弹性伸缩
5. 多分片（一般为5），3副本

Squirrel技术架构

应用方	申请集群		
Squirrel	申请docker实例		构建集群 交付
Hulk	申请物理机		生成docker实例
基础设施			交付物理机 加入hulk资源池

客户端路由

1. 多种客户端路由策略：
    1. master-only
    2. slave-only
    3. IDC就近路由
2. 启动访问zookeeper
3. 集成CAT



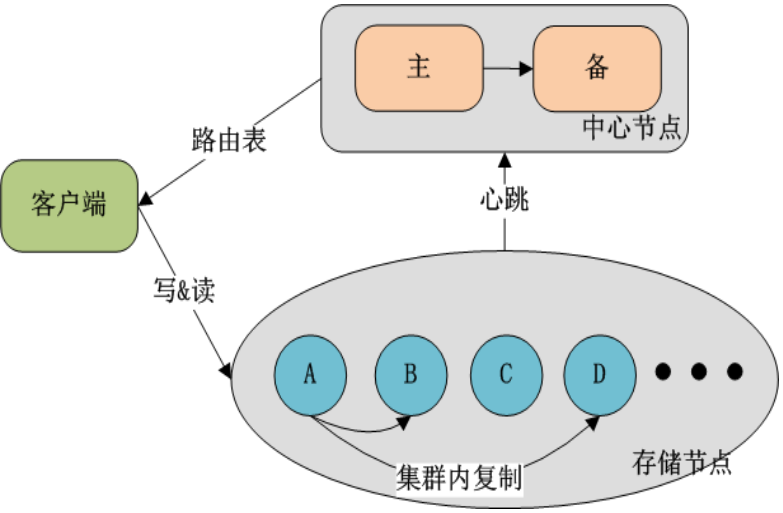
三大存储发展现状

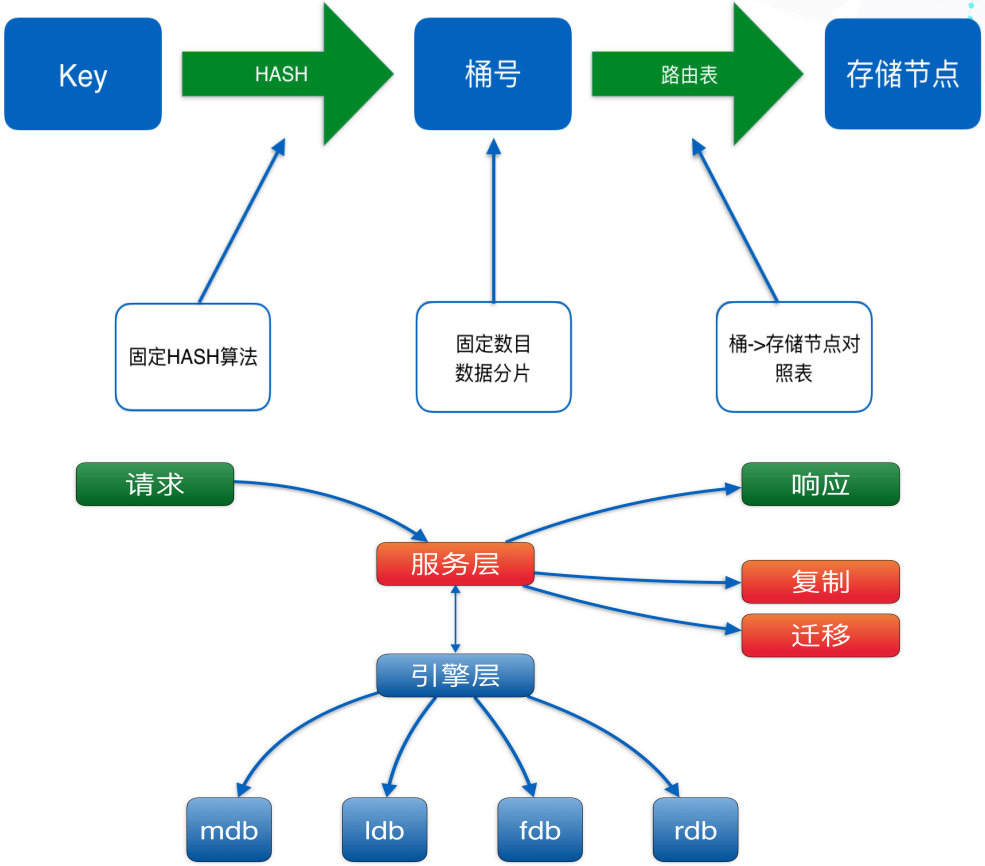
之Cellar：　分布式存储，有容乃大

Cellar服务现状

1.14年初，引入阿里Tair作为NoSQL存储
2.14年底，大范围应用，并对Tair修修补补，积累领域问题
3.16年初，基于开源版本研发新一代KV存储系统Cellar
4.日请求量达数万亿级，美团点评最大NoSQL存储

Tair起源-架构


主
备
中心节点
路由表
客户端
心跳
写&读
A
B
C
D
集群内复制
存储节点


Key
HASH
桶号
路由表
存储节点
固定HASH算法
固定数目
数据分片
桶->存储节点对
照表
请求
响应
服务层
复制
迁移
引擎层
mdb
ldb
fdb
rdb

Tair起源-架构

架构升级

可用性优化

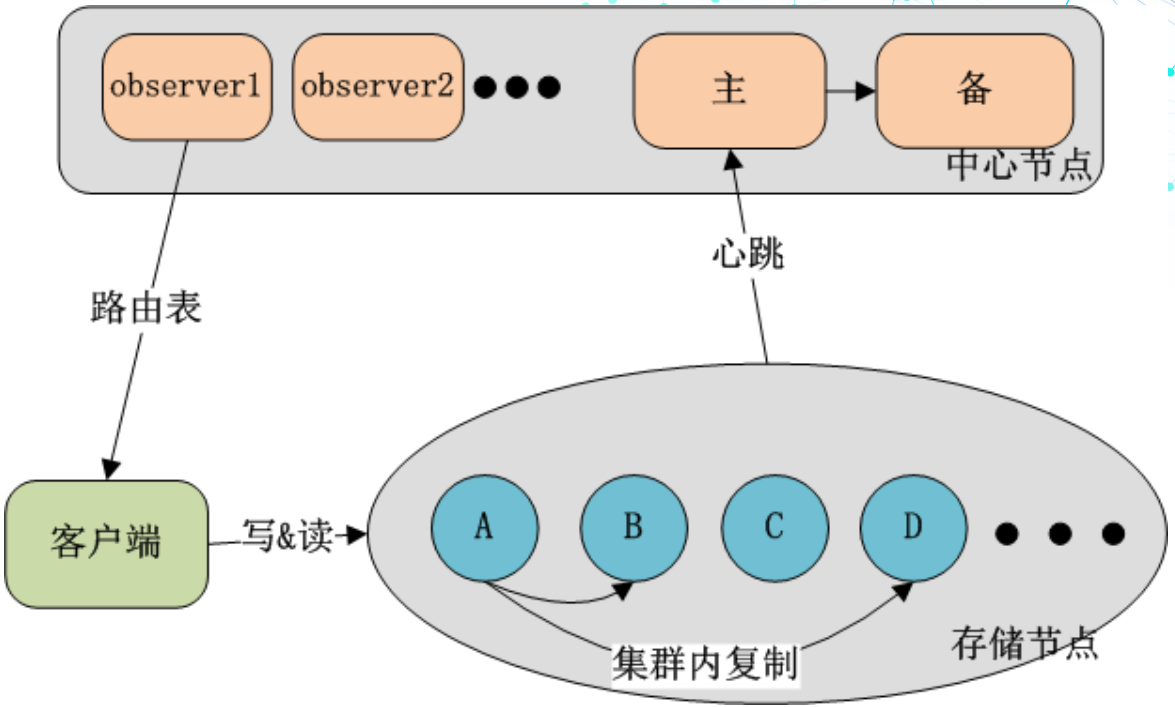
性能优化

可运维性

平台建设

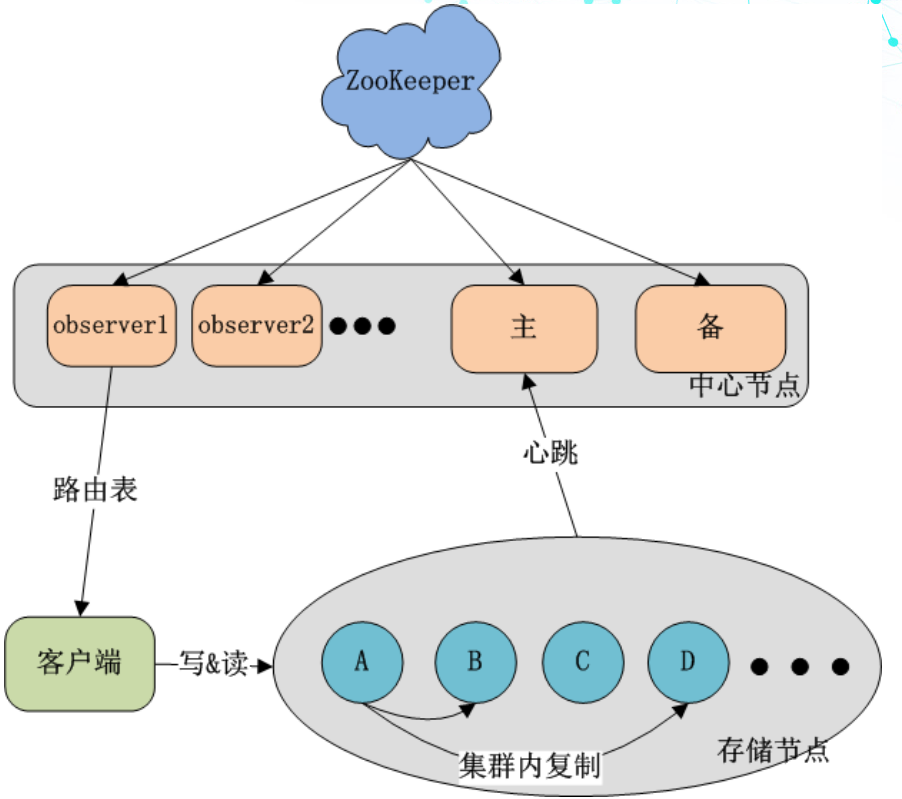
Cellar改进

- 可扩展性：
  - 路由查询能力
  - 可线性扩展
- 隔离性：
  - 客户端与中心节点
  - 完全隔离



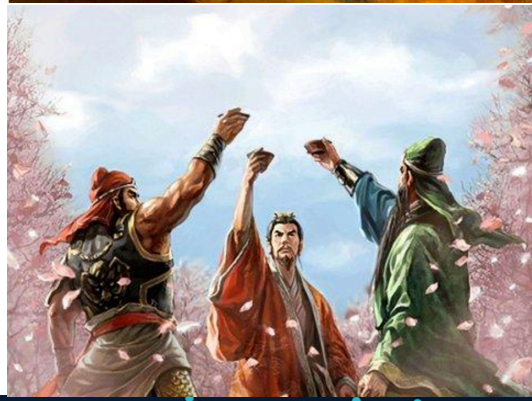
Cellar改进

- 一致性：
  - 主备强一致
  - observer同步强一致



02

天下大势，
分久必合，
合久必分，
烽火狼烟，
战争四起！



存储的竞争与融合

MySQL特性

1.关系型数据库，支持二级索引

2.B+ Tree结构，支持范围扫描

3.支持join和聚合查询

4.二进制日志，支持point-in-time recovery

5.结构化+行存储，在线DDL是痛点（pt-osc与gh-ost结合）

6.支持事务和ACID，可配置的一致性策略，金融领域必备

7.读多写少的应用场景，结合新硬件，响应时间降低到1ms左右

8.写入能力无法线性扩展；从库同步单线程造成延迟

Squirrel特性

1.丰富的数据结构，易用

2.备份AOF和RDB，4.0开始持久化得到增强

3.对响应时间要求极高，数据规模可预见

4.热点数据比较集中，数据规模中等（比如300G以内）

5.单线程

   1.发号器

   2.不存在冲突，与lua结合实现秒杀

   3.不适合大批量导入数据

Cellar特性

1. 存储缓存二合一
2. 节点升级，缓存数据不需要重新预热
3. 线性扩容
4. 自动负载均衡
5. 数据规模：大型或超大型-达到几十T的规模
6. 适合存储小图片、大量的文字描述等数据
7. 不适合：
   发号器
   对单一key频繁更新
   热点数据不均匀
   大量查询不存在的key

场景之争

数据能自动过期吗?

能在线加字段吗?

能水平扩容吗?

删库跑路了能找回吗?

有权限控制吗?

主从数据可能会不一致吗?

从库有延迟吗?

故障可能导致脑裂吗?

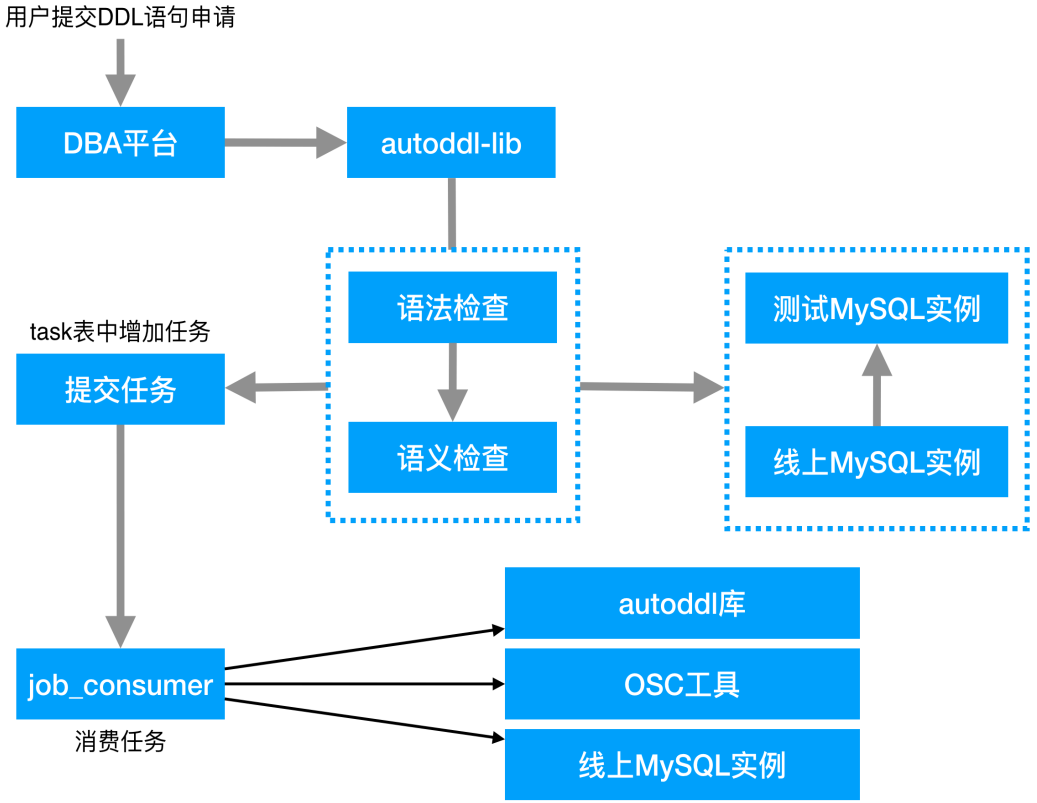
能自动读写分离吗?

tradeoffs are hard

工具链和平台建设

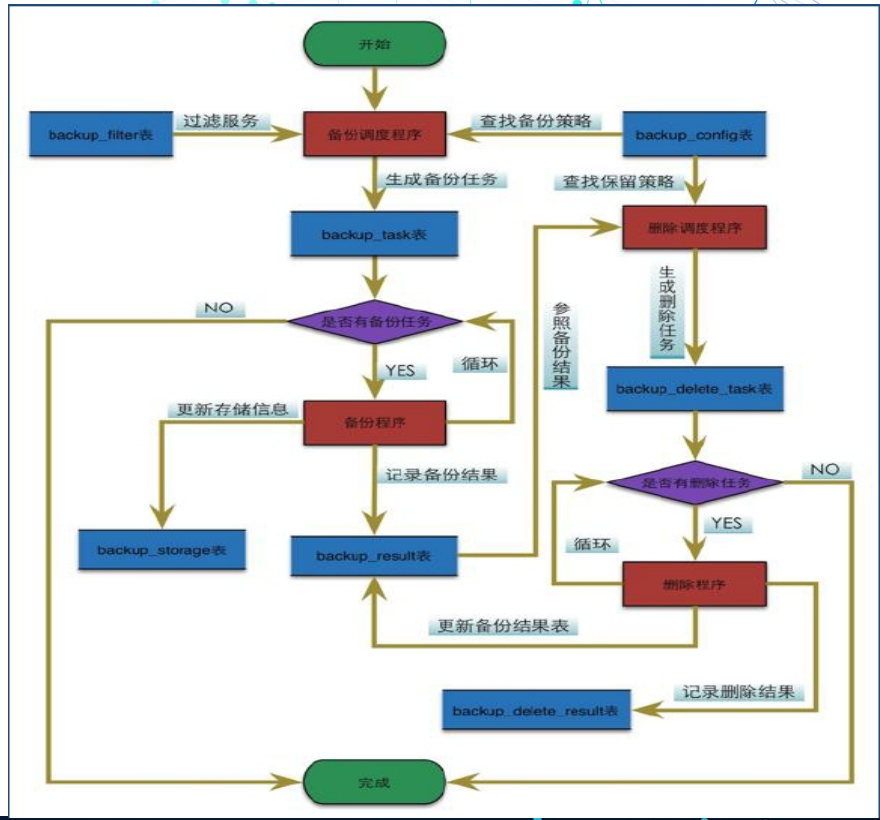
自助DDL、DML

1.语法检测和语义检测
2.单个集群串行调度
3.可选pt-osc或gh-ost
4.可按集群灰度不同版本工具
5.业务自定义高峰时间段
6.DML执行前进行备份



银河备份恢复系统

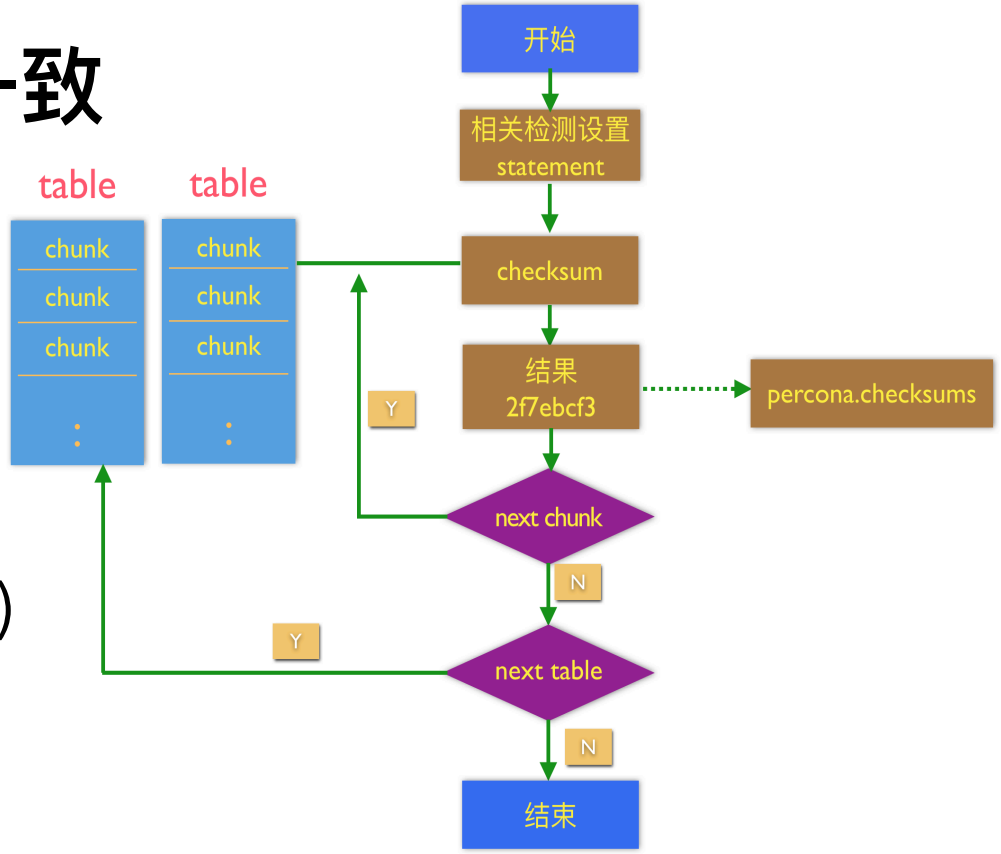
1.基于xtrabackup工具备份
2.本地备份+远程云存储
3.改造blackhole引擎备份binlog
4.对备份数据进行有效性测试
5.对binlog能否接上进行监控
6.基于flashback工具快速恢复
7.尝试binlog-server做HA+备份



数据一致性校验

多种原因引起数据不一致

1.主从切换丢数据
2.主库down机恢复为从库
3.备份恢复软件bug
4.升级版本导致（datetime）
5.主从同步bug（5.7的多线程）



报警自动分析与处理

[报警分析结果报告][p2][[redacted]][Slow_queries][2017-03-17 15:18:30]
[报警处理方案]
[1: kill file sort的连接(5s,1min)]
[2: kill tmp table的连接(5s,1min)]
[3: kill正在运行的select连接(5s,1min)]

[报警分析结果报告][p0][[redacted]01][Threads_running][2017-03-12 07:35:00]
[报警处理方案]
[1: kill正在运行的select连接(5s,1min)]
[2: kill file sort的连接(5s,1min)]
[[3: kill sleep的连接30s,1min]]
[4: kill tmp table的连接(5s,1min)]
[5: kill Creating sort的连接(5s,1min)]
[6: kill Copying to group table的连接(5s,1min)]
[7: kill converting HEAP to MyISAM 的连接(5s,1min)]

[报警分析结果报告][p0][[redacted]04][load.1minPerCPU][2017-03-14 16:03:00]
[报警处理方案]
[1: kill file sort的连接(5s,1min)]
[2: kill tmp table的连接(5s,1min)]
[3: kill正在运行的select连接(5s,1min)]
[4: kill Creating sort的连接(5s,1min)]
[5: kill Copying to group table的连接(5s,1min)]
[6: kill converting HEAP to MyISAM 的连接(5s,1min)]
[7: kill 处于Waiting for 的连接(5s,1min)]
[8: kill sleep的连接(30s,1min)]

美团点评DBA

[报警分析结果报告][p0][[redacted]01][Conn_used_ratio][2017-03-20 20:08:00]
[报警处理方案]
[1: kill sleep的连接(30s,1min)]

美团点评DBA

[预处理执行成功][[redacted]01][Conn_used_ratio][2017-03-20 20:09:26]

美团点评DBA

[有效处理执行成功][[redacted]01][Conn_used_ratio][2017-03-20 20:10:32]

执行脚本: hid_sleep_kill.sh
执行描述: kill sleep的连接(30s,1min)
执行命令:

```
pt-kill[redacted]idle-time=30s --run-time=1m  --print --kill --victims all --interval 0.1
pt-kill[redacted]idle-time=30s --run-time=1m  --print --kill --victims all --interval 0.1
```

执行细节:

```
预处理前期check都已成功,开始进行预处理.

# A software update is available:
#   * Percona Toolkit 2.2.6 has a possible security issue (CVE-2014-2029) upgrade is recommended. The current version for Percona::Toolkit is 2.2.20.

# 2017-02-03T10:01:15 KILL 3843980 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843941 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843931 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843842 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843730 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843595 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843460 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843450 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843431 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843367 (Sleep 71 sec) NULL
# 2017-02-03T10:01:15 KILL 3843079 (Sleep 71 sec) NULL
```

优化、诊断、跟踪

开源工具	内部平台	用途说明
pt-query-digest	慢查询分析平台	慢sql统计分析、跟踪
pt-stalk-rainGauge	雨量计系统	rootCause跟踪分析，灵活设置触发规则
pt-kill+daemontools	过载保护系统	慢查询、大事务自动杀死
pt-upgrade	升级导流、对比测试	5.6升级到5.7，percona迁移到mariadb，或pxc，或mgr，对响应时间、返回值对比
pt-config-diff	配置对比改写系统	主从配置对比；内存值与my.cnf对比；多个从库之间对比（group_concat_max_len悲剧）

监控平台

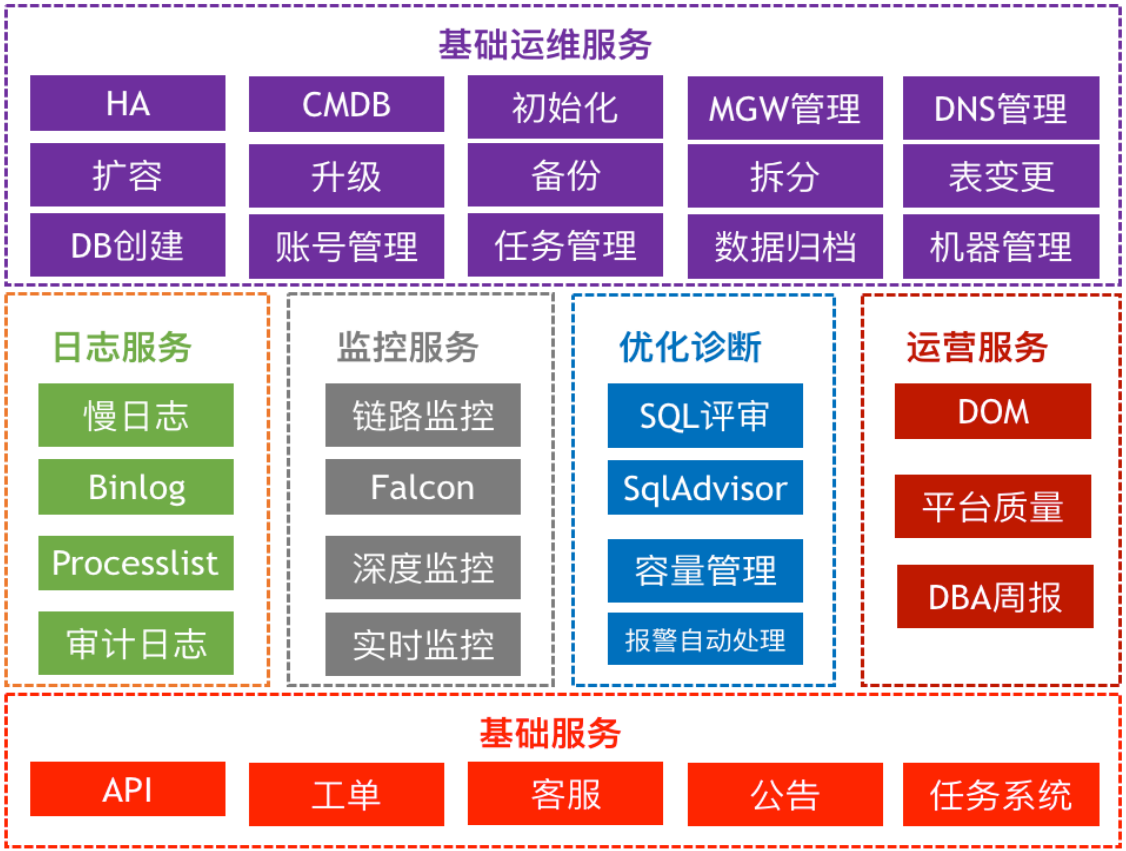
1.服务器监控：MT-falcon
    1.基于小米开源open-falcon改造
    2.采集服务器端指标并报警
2.端到端监控：CAT（Central Application Tracking）
    1.基于Java开发，在业务代码中埋点上报
    2.以客户端视角看服务

海豚自助平台2.0

海豚自助平台2.0


用户
DBA
运维
测试
开发
产品


平台
业务管理
业务监控


基础运维服务
HA
CMDB
初始化
MGW管理
DNS管理
扩容
升级
备份
拆分
表变更
DB创建
账号管理
任务管理
数据归档
机器管理
日志服务
慢日志
Binlog
Processlist
审计日志
监控服务
链路监控
Falcon
深度监控
实时监控
优化诊断
SQL评审
SqlAdvisor
容量管理
报警自动处理
运营服务
DOM
平台质量
DBA周报
基础服务
API
工单
客服
公告
任务系统


标准化
工具化
自助化
接口化
智能化

Cellar管理平台2.0

1.底层基于ansible-playbook+shell开发
2.上层调用playbook提供的API，命令行与web入口共享
3.解决了重复开发问题
4.对开发屏蔽了运维细节
5.很好的应对了业务的高速膨胀
6.升级、降级、缩容、扩容、迁移，全部实现自助化
7.通过运维平台建设，95%的操作实现自助化

用数据说话

抓主要矛盾，定向爆破

1.按报警类型统计
2.按集群统计
3.按负责人统计
4.按时间聚合

服务组不健康度	收到OK短信条数	name	dba_user
5	7	[redacted]oordinatehis	[redacted]
4	4	[redacted]open	[redacted]
4	4	[redacted]_stat	[redacted]
4	7	[redacted]est	[redacted]
4	8	[redacted]officepr	[redacted]
4	6	[redacted]rch	[redacted]
4	4	[redacted]_data	[redacted]
3	4	[redacted]mtsi	[redacted]
3	3	[redacted]ight	[redacted]
3	5	peg[redacted]	[redacted]

alarm_info	times
Seconds_Behind_Master	[redacted]
load.1minPerCPU	[redacted]
df.bytes.free.percent	[redacted]
Conn_used_ratio	[redacted]
Slave_IO_Running	[redacted]
Threads_running	[redacted]
Slave_SQL_Running	[redacted]
Uptime	[redacted]
mysql_alive_local	[redacted]
net.if.out.Mbps	[redacted]
mem.swapused.percent	[redacted]
icmp.ping.alive	[redacted]
Innodb_deadlocks	[redacted]
cpu.idle	[redacted]
cpu.freq_avg	[redacted]

select count(distinct ceil(unix_timestamp(alarm_time)/1800)) 服务组不健康度 from 报警表

多维度衡量服务质量

1.慢查询占比
2.绿帽子和出轨占比
3.服务化程度衡量
4.数据库容量预警
5.响应时间95线、99线
6.服务器低利用率统计
7.topN统计
8.红黑榜

总结

MySQL

NoSQL

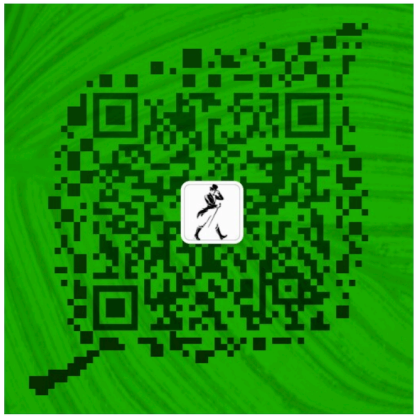
Not Only MySQL

MySQL is a better NoSQL

MGR，maybe the game changer?

我们在招聘DBA+SRE

- 为gh-ost贡献issue+patch
- 为mha贡献issue+patch
- 开源的dbproxy、sqladvisor、zebra，star数量超过1000+
- Cellar计划开源（基础架构部开发）
- 负责外卖、配送、金融、酒旅、猫眼各业务线
- 邮箱：zhaoyinggang@meituan.com



Thanks

关注开源数据库论坛