TDSQL在金融场景中的实践

腾讯云

harlylei(雷海林) 专家工程师

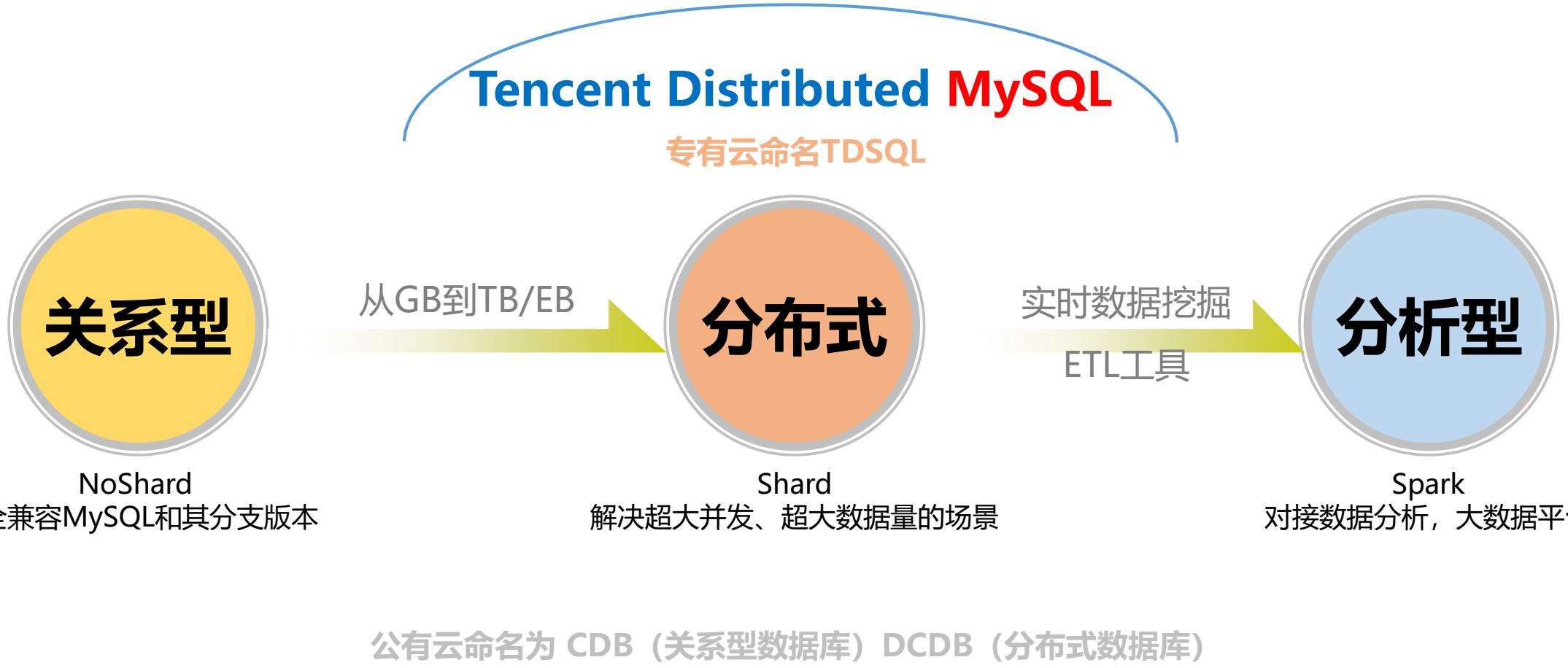
目录

1 TDSQL简介

定位、合作伙伴

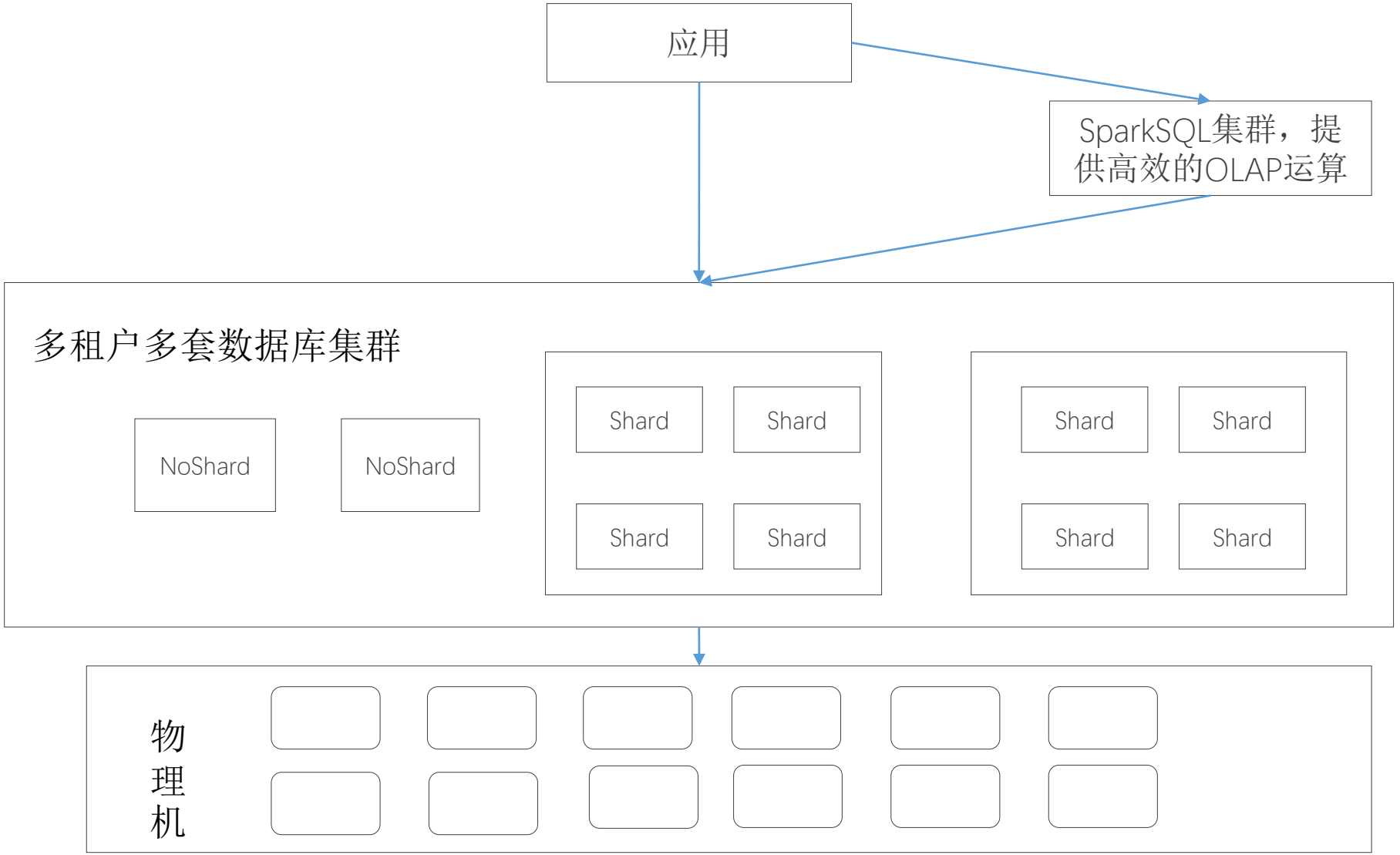
TDSQL（Tencent Distributed MySQL）是腾讯针对金融场景推出的强一致性，分布式数据库集群解决方案。

介绍



公有云命名为 CDB（关系型数据库）DCDB（分布式数据库）

灵活的多租户集群

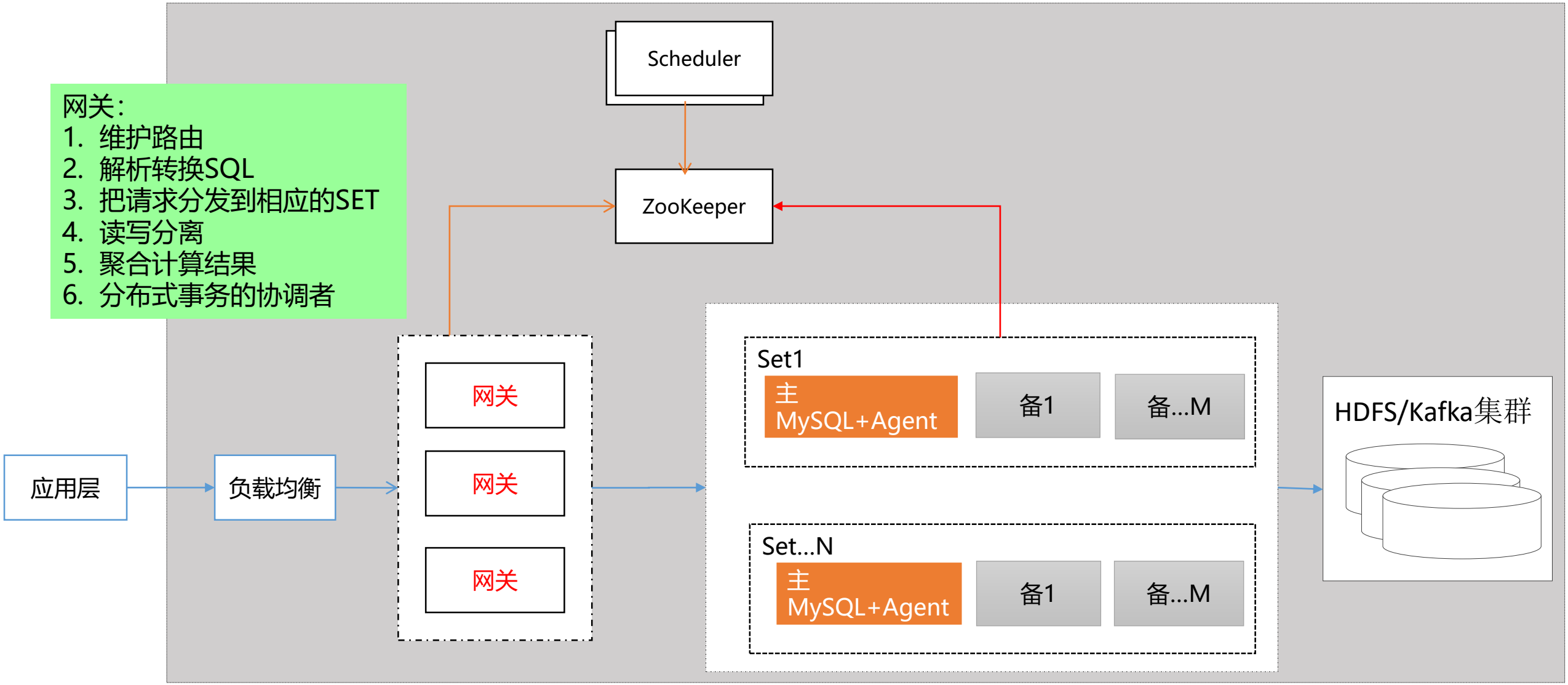

应用
SparkSQL集群，提供高效的OLAP运算
多租户多套数据库集群
NoShard
NoShard
Shard
Shard
Shard
Shard
Shard
Shard
Shard
Shard
物理机

2 核心特性

高一致性、高可用性、数据保护

水平扩展、高性能、自动化运营体系

TDSQL核心架构


网关：
1. 维护路由
2. 解析转换SQL
3. 把请求分发到相应的SET
4. 读写分离
5. 聚合计算结果
6. 分布式事务的协调者
Scheduler
ZooKeeper
网关
网关
网关
应用层
负载均衡
Set1
主
MySQL+Agent
备1
备...M
Set...N
主
MySQL+Agent
备1
备...M
HDFS/Kafka集群

数据库引擎

默认支持

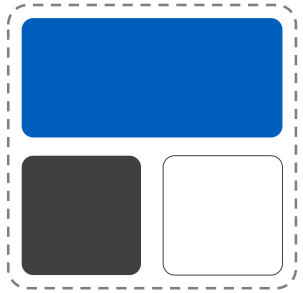
MariaDB

10.1　10.0

兼容MySQL 5.6

Percona/MySQL

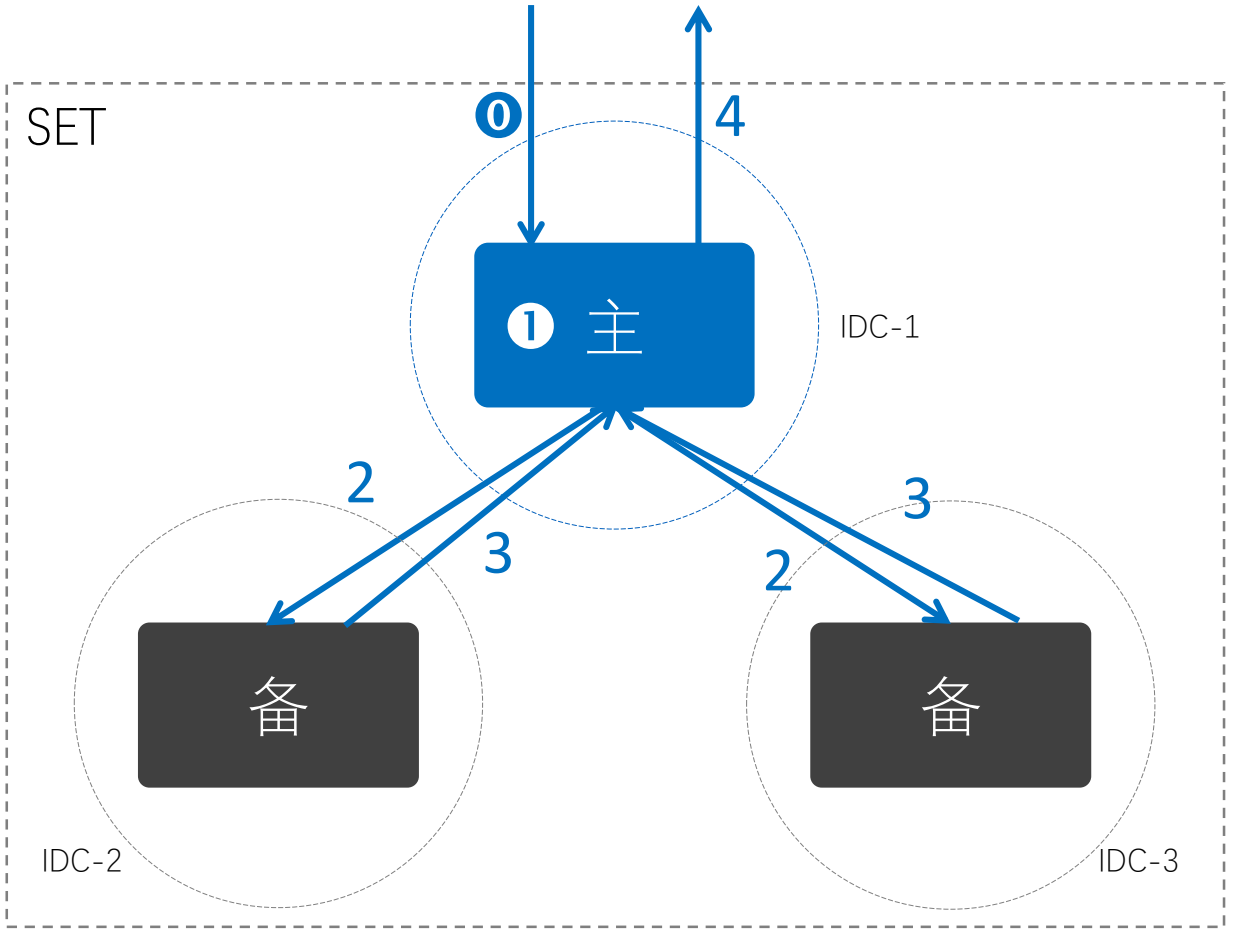
5.7



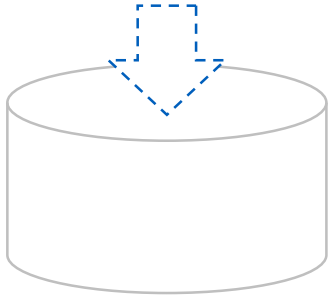
复制

Replica

强同步一致性协议

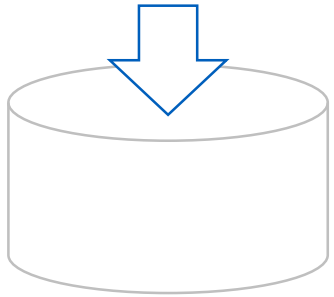


主备数据复制方式



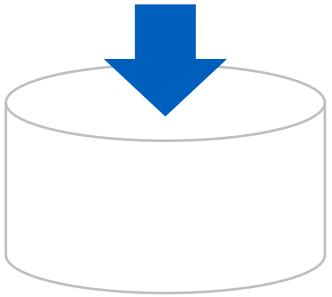
异步复制

Async replication



半同步复制

Semi-Sync replication



强同步复制

Sync replication

半同步复制的不足

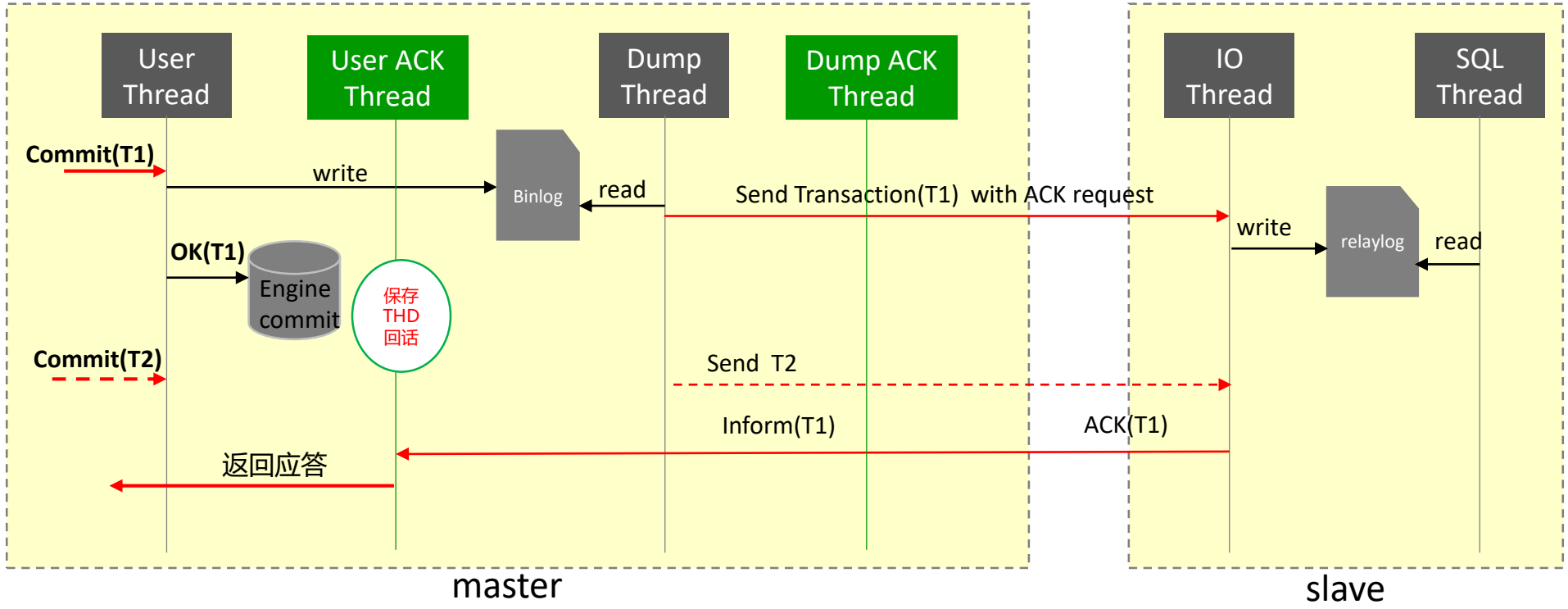
1.超时后蜕化成异步，金融场景不合适

2.跨IDC的情况下性能不乐观(2-3ms)

update_non_index.lua

复制性能对比（跨IDC）	TPS
异步	60,000
5.6半同步	5000
5.7半同步	30000

用户线程异步化



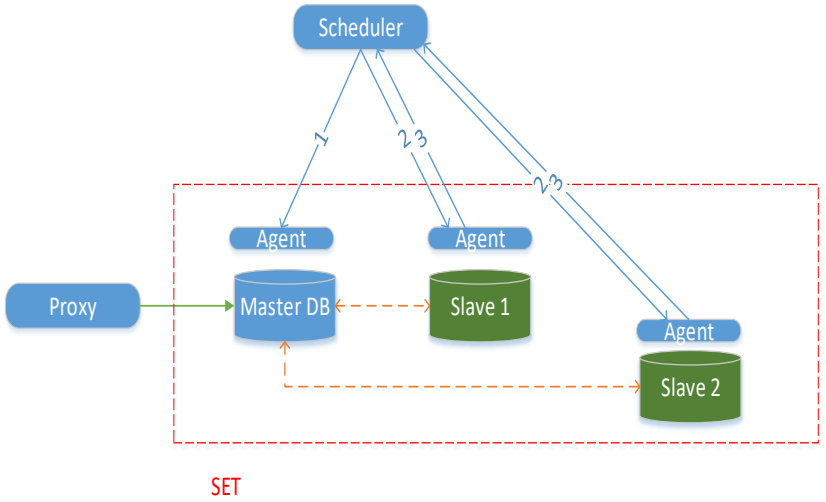
复制性能对比（跨IDC）	TPS
异步	60,000
5.6半同步	5000
5.7半同步	30000
TDSQL强同步	60000

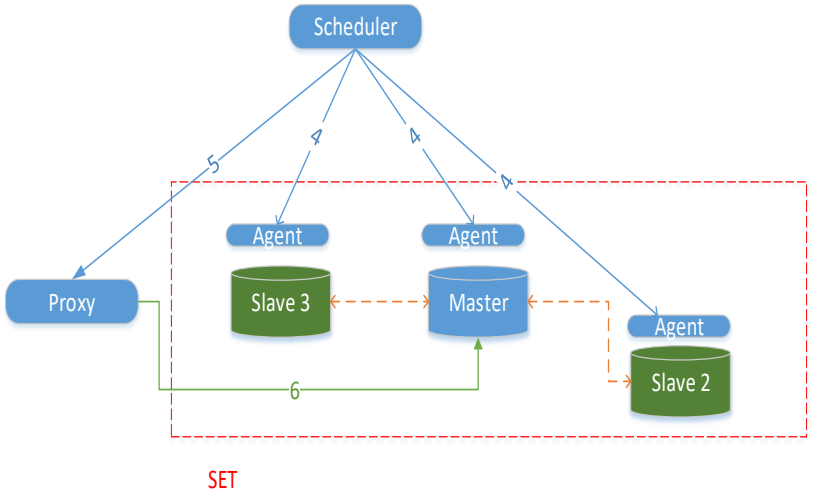
高一致性容灾 — 如何保证没有脏数据

原则：

1、主机可读可写，备机只读，备机可以开放给业务查询使用

2、任何时刻同一个SET不能有两个主机





1、主DB降级为备机

2、参与选举的备机上报最新的binlog点

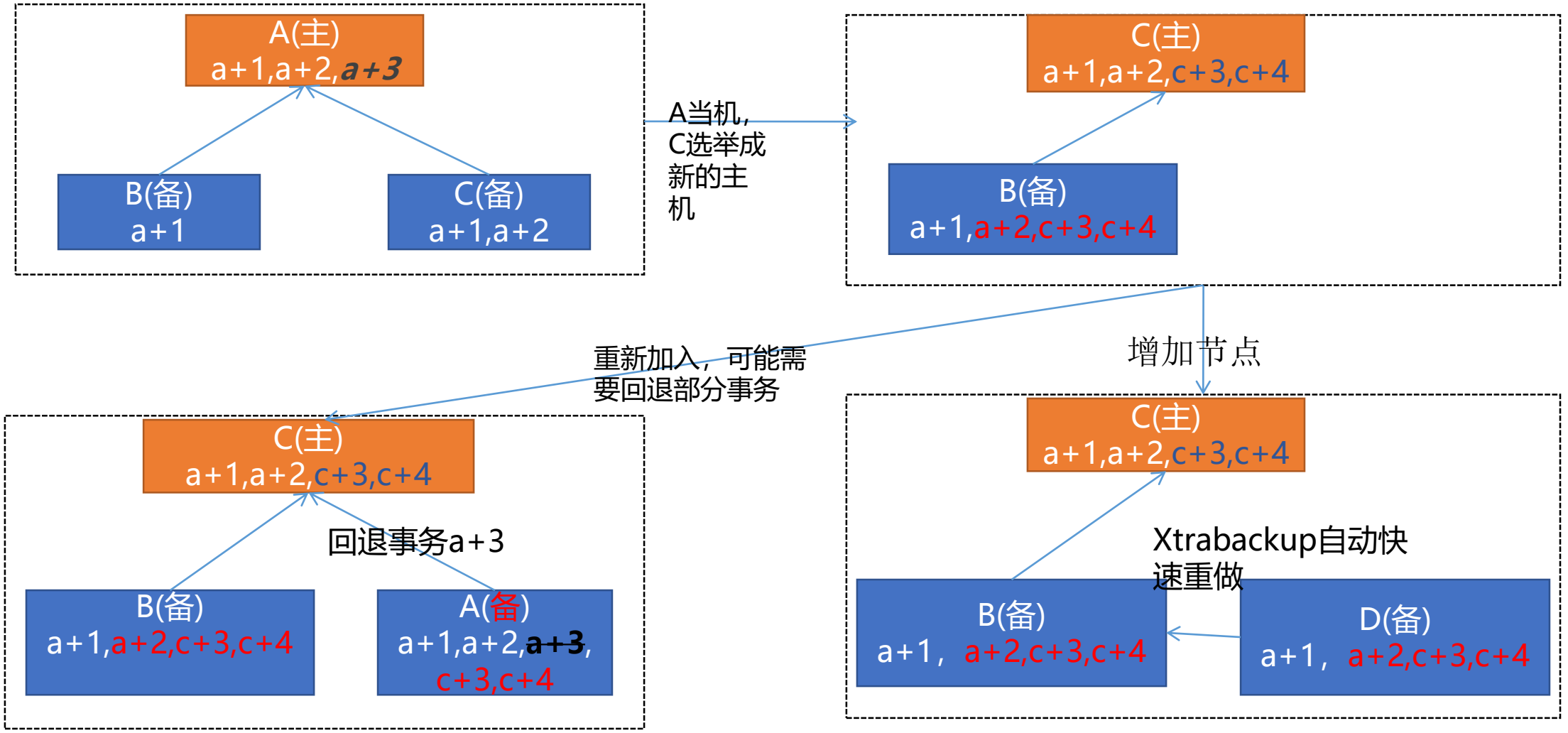
3、scheduler收到binlog点之后，选择出binlog最大的节点

4、重建主备关系

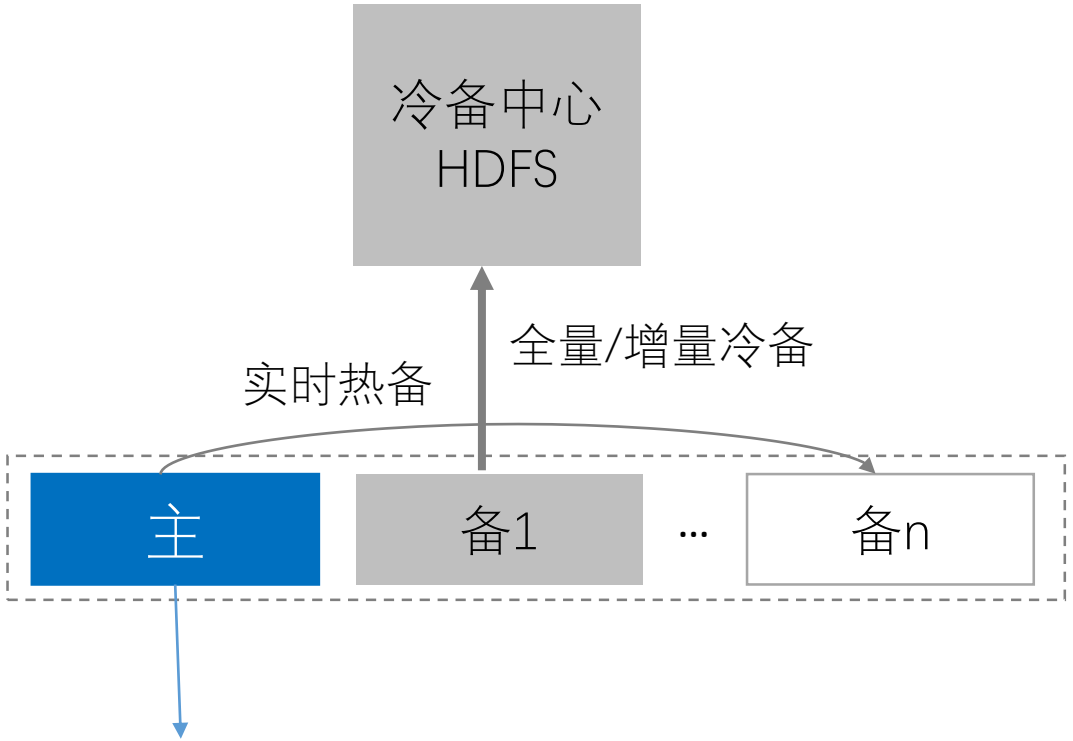
5、修改路由

6、请求发给新的主机

数据高可用性的保障机制（恢复）



数据备份和恢复



1.数据闪回功能
2.数据库表延迟删除

物理备份

1.利用xtrabackup备份物理文件。
2.每天凌晨或者指定时间点备份一次。
3.压缩备份到HDFS。
4.支持增量备份
5.支持通过管理台直接触发备份

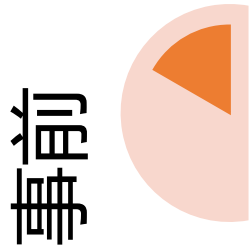
逻辑备份

1.利用mydumper备份工具。
2.每天备份一次。
3.压缩备份到HDFS。
4.每个库、表结构、表数据独立备份，恢复时可以单独恢复。
5.支持通过管理台直接触发备份

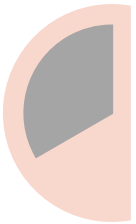
binlog备份

1.实时备份binlog到hdfs。lz4压缩。

全面安全防护



- IP白名单
- SSL连接加密
- SQL防火墙



- 文件透明加密
- 网络隔离
- 运维安全保障

- 数据库审计
- 操作日志审计
- 服务器审计

读写分离

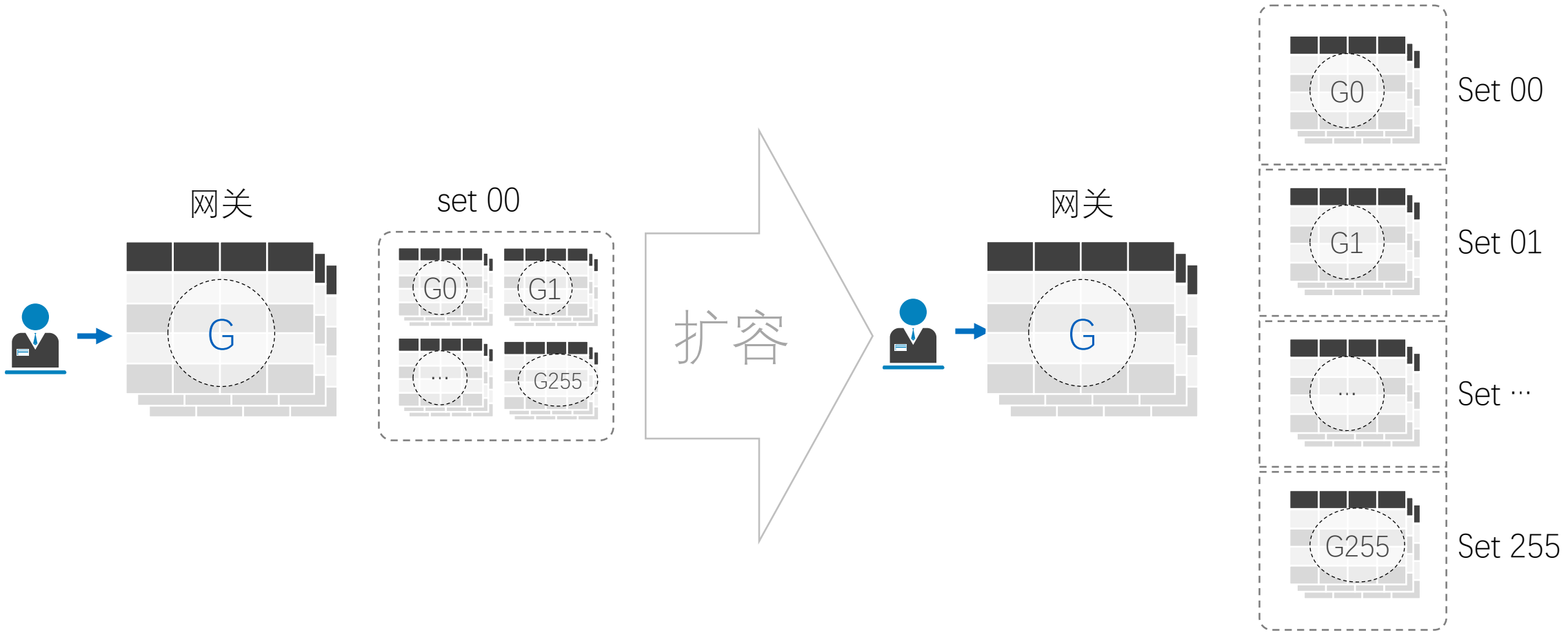
- 基于数据库账号的读写分离
- 自动分析事务进行读写分离
- 基于Hint的读写分离

```
//主机读//
select * from emp order by sal, deptno desc;
//从机读//
/*slave*/ select * from emp order by sal, deptno
```

3 分布式实践

扩展性、分布式事务

Shard水平扩容



三种表

- **Shard表**
  - create table account(user int , payamt int, c char(20) ,PRIMARY KEY (user)) **shardkey=**user;
  - create table bill(user int , billno int, c char(20) ,PRIMARY KEY (user)) **shardkey=** user;

  - create table dummytable(seqno int , c char(20) ,PRIMARY KEY (seqno)) **shardkey=** seqno;

- **NoShard表**，如一些简单的配置表
  - create table noshard_table (a int, b int key, PRIMARY KEY (a));

- **广播表**，支持全局广播
  - create table global_table (a int, b int key, PRIMARY KEY (a)) shardkey=noshardkey_allset;

SQL支持-基本兼容所有SQL

- group by, order by
- max, min, sum, avg等聚合函数
- distinct, count
- Join
- Transaction (分布式事务)

分布式事务

XA

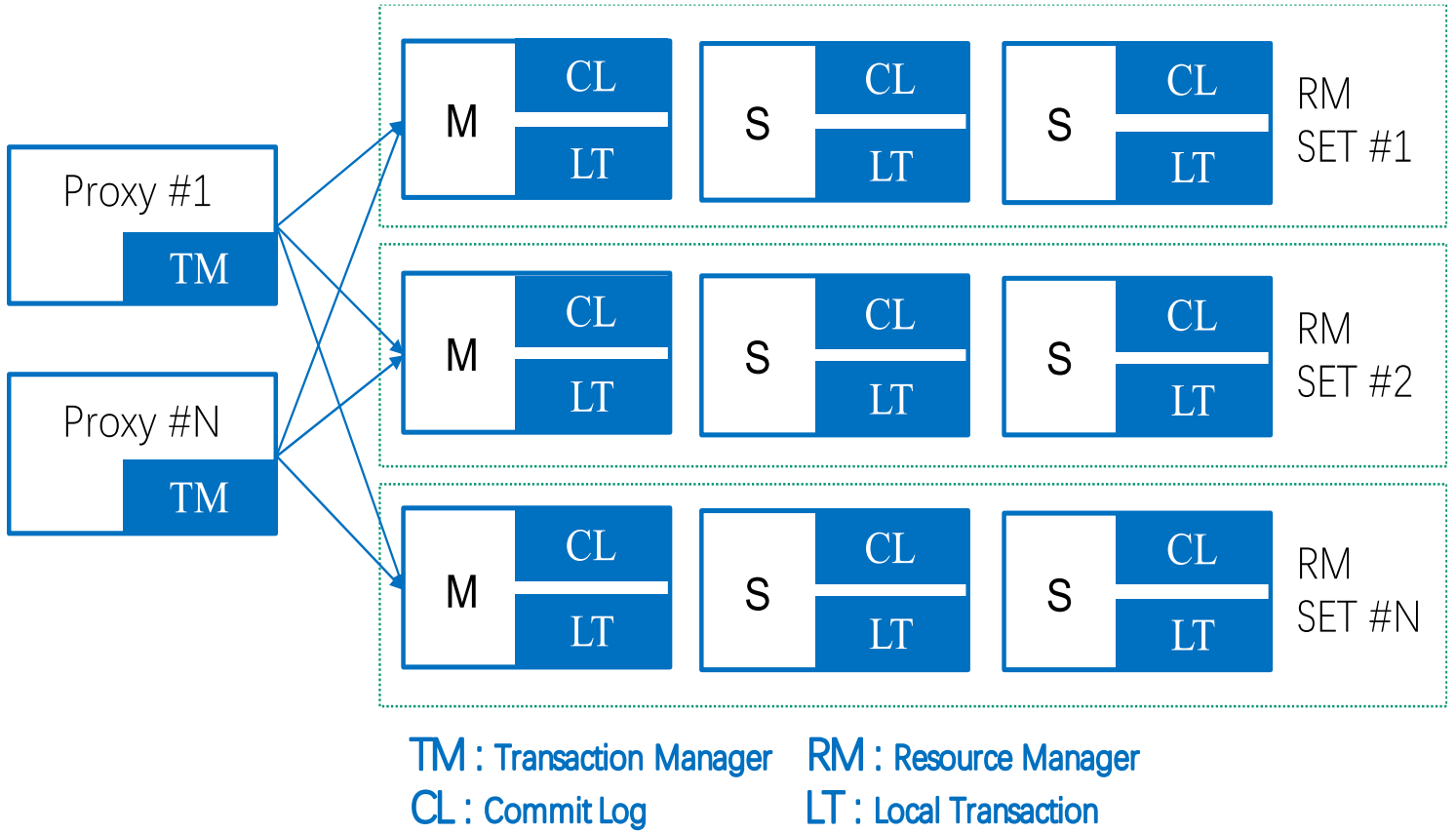
分布式事务

完全去中心化、性能线性增长

健壮的异常处理

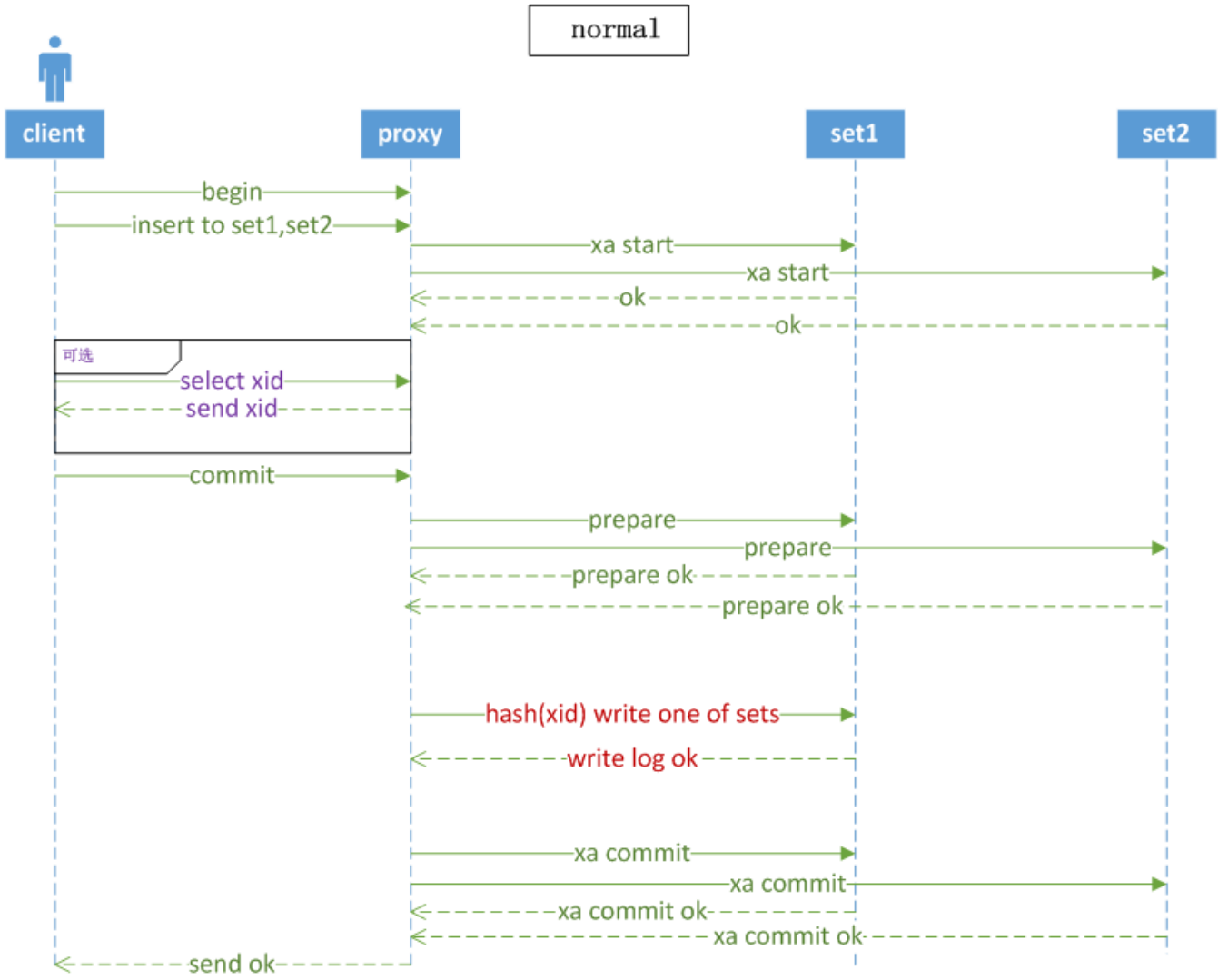
全局死锁检测机制

TPCC标准验证



分布式事务

- Prepare 超时或者失败
- Commit log写失败
- Commit log写超时
- Commit超时或者失败
- 异常的总结



选型推荐

中小规模

- 数据量：< 4T
- tps：< 8K/s
- 通用性 > 容量伸缩

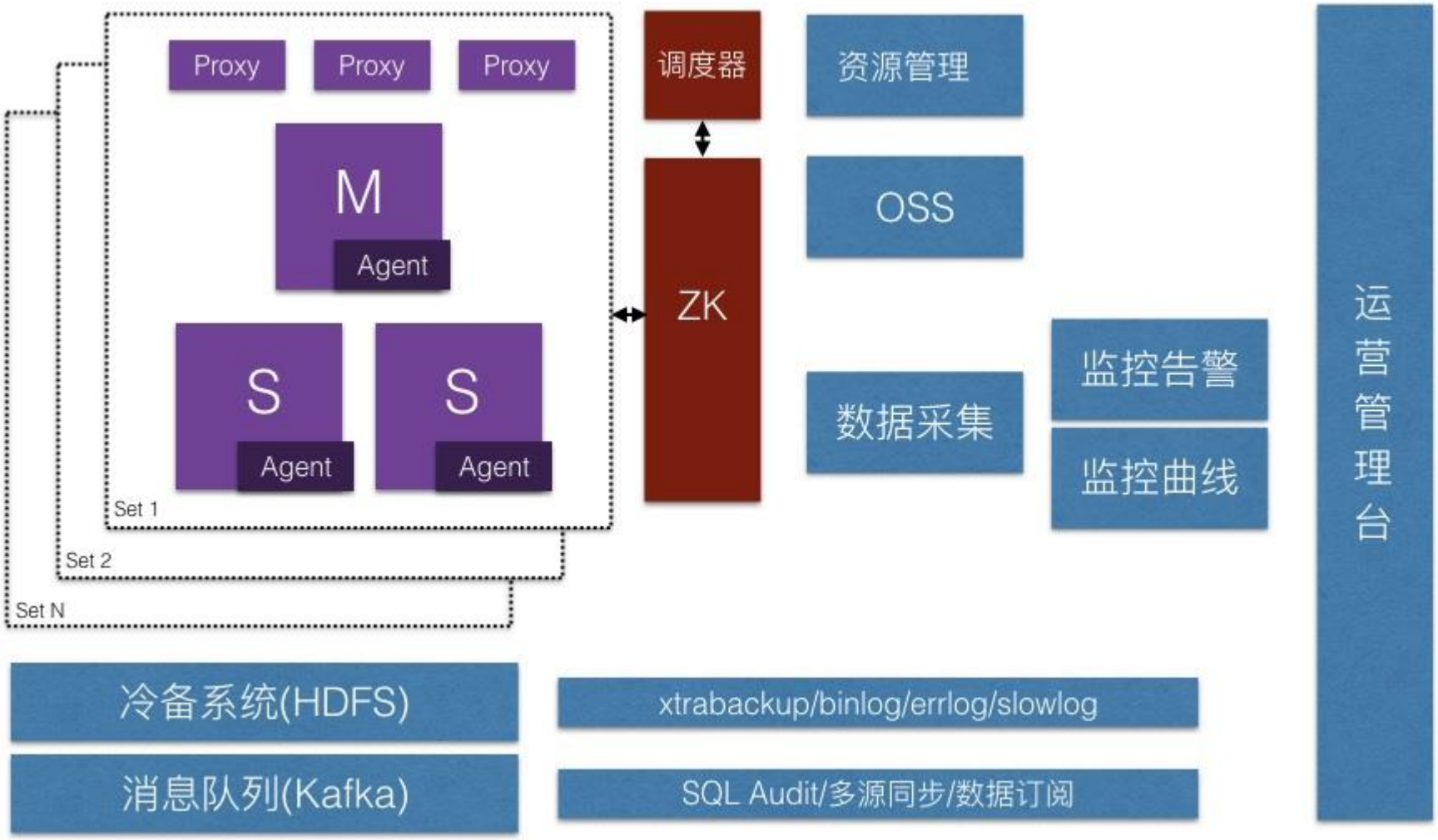
巨大规模

- 数据量：> 4T
- tps：> 8K/s
- 通用性 < 容量伸缩

NoShard

Shard

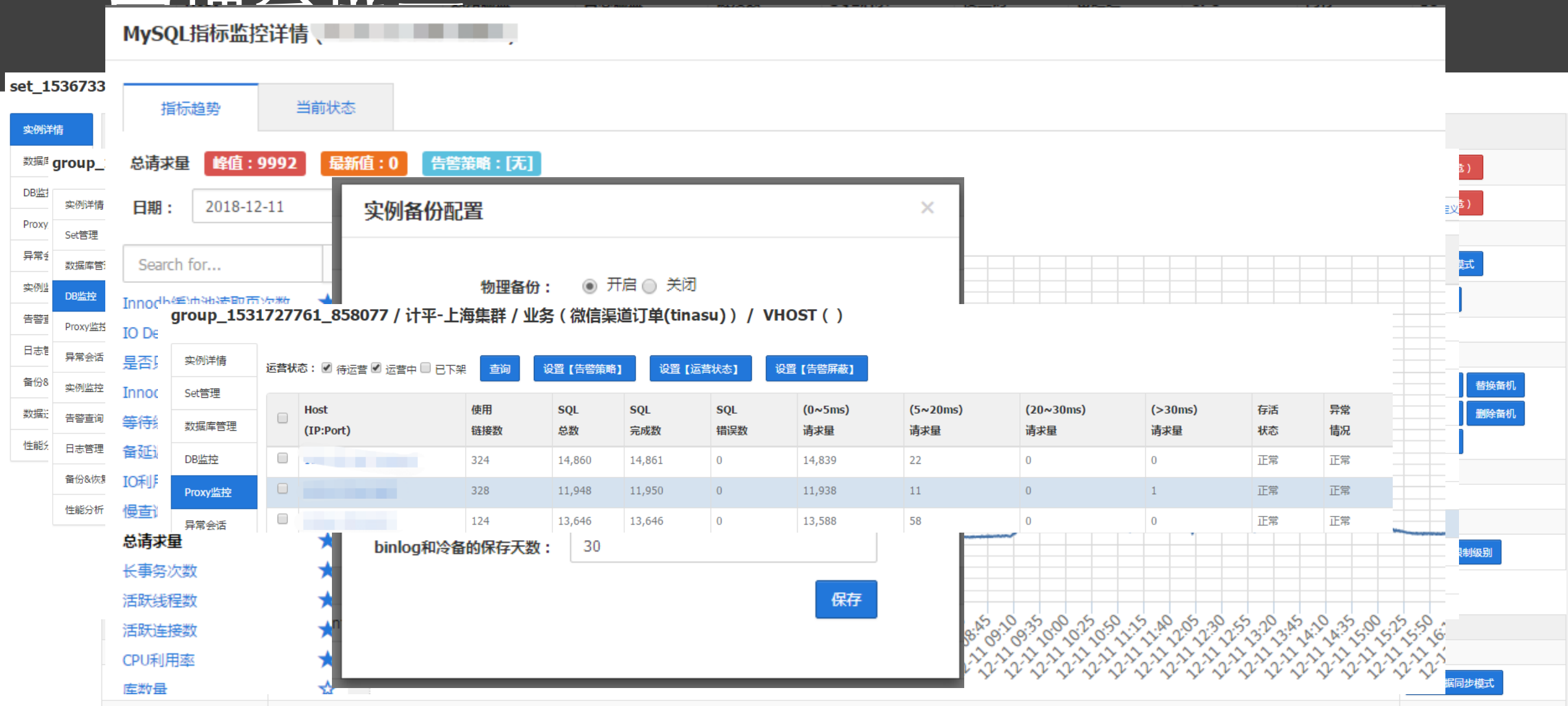
TDSQL整体视图


Proxy
Proxy
Proxy
M
Agent
S
Agent
S
Agent
Set 1
Set 2
Set N
调度器
ZK
资源管理
OSS
数据采集
监控告警
监控曲线
运营管理台
冷备系统(HDFS)
消息队列(Kafka)
xtrabackup/binlog/errlog/slowlog
SQL Audit/多源同步/数据订阅

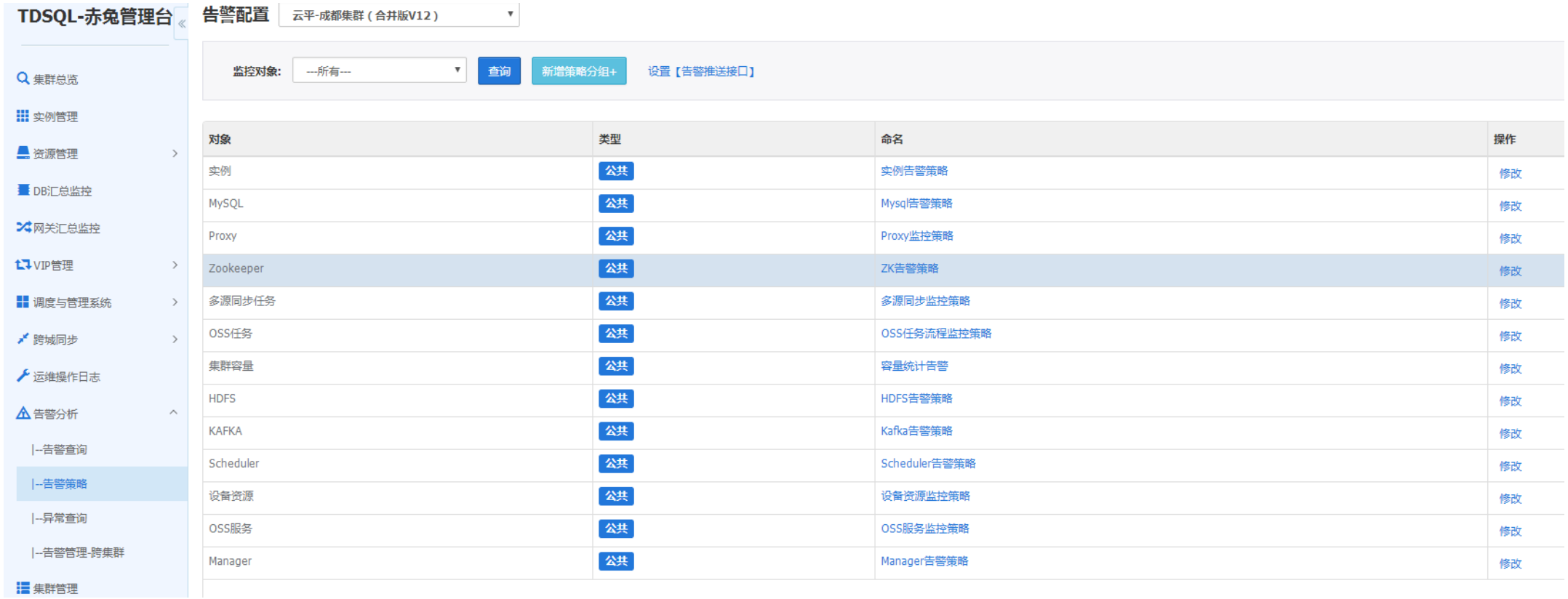
赤兔管理平台


TDSQL运营管理
请输入实例名、业务名、
全局搜索
友情链接:
云平九寨系统
云平监控barad
[注销] 潘安群（aaronp
集群总览
运营操作
运营监控
运营告警
业务运营
接口管理
系统管理
|--用户管理
|--角色权限
|--集群配置
|--操作日志
|--系统版本
使用量
总容量
使用率(%)
实例异常
Mysql异常
Proxy异常
0
0
0
计平-理财通集群
>进入
使用量
总容量
使用率(%)
实例异常
Mysql异常
Proxy异常
0
39
1
计平-深圳集群2
>进入
使用量
T
总容量
使用率(%)
95.39
实例异常
Mysql异常
Proxy异常
0
29
9
计平-深圳集群
>进入
使用量
T
总容量
使用率(%)
实例异常
Mysql异常
Proxy异常
使用量
G
总容量
使用率(%)
实例异常
Mysql异常
Proxy异常
使用量
T
总容量
使用率(%)
实例异常
Mysql异常
Proxy异常

管理台展示

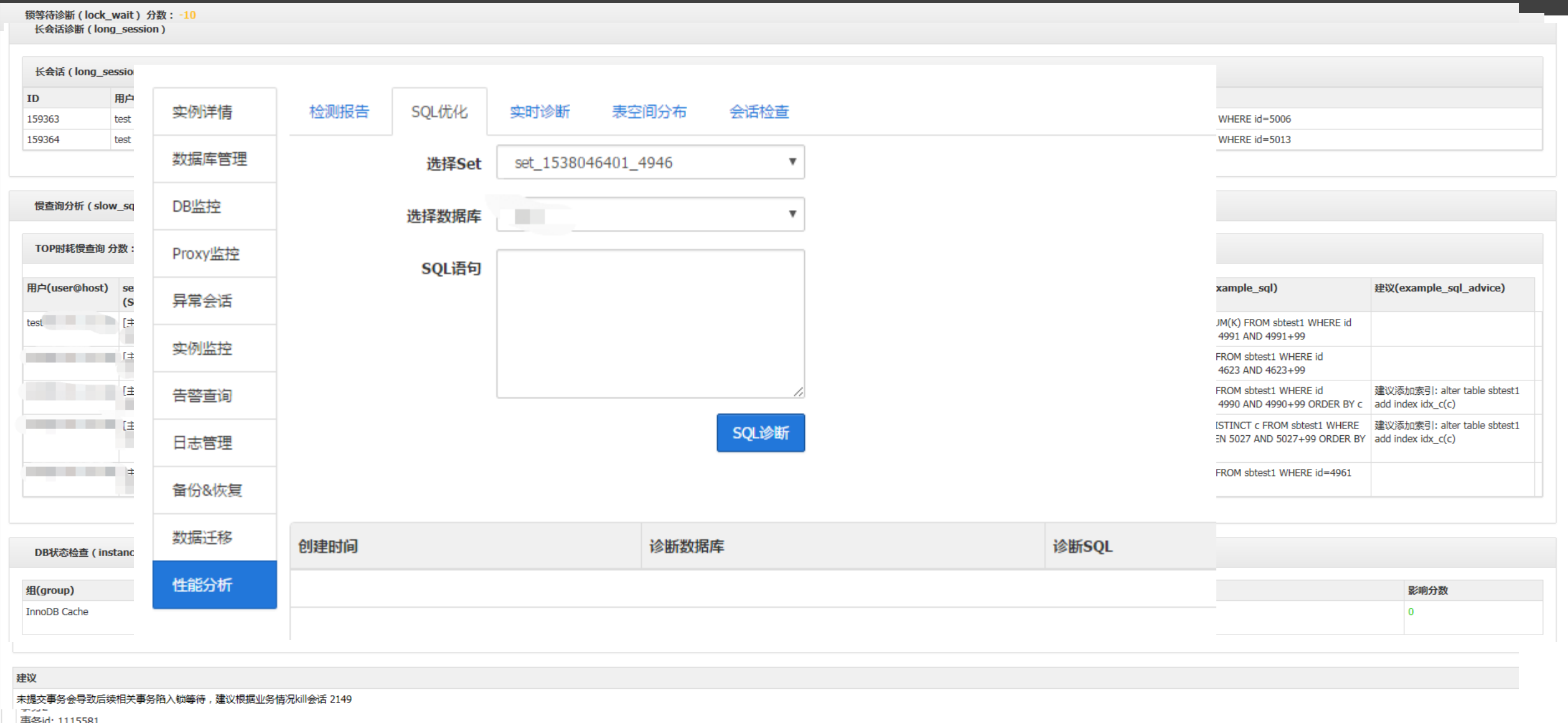

MySQL指标监控详情
指标趋势
当前状态
总请求量
峰值：9992
最新值：0
告警策略：[无]
日期：
2018-12-11
Search for...
实例详情
Set管理
数据库管理
DB监控
Proxy监控
异常会话
实例监控
告警查询
日志管理
备份&恢复
性能分析
实例备份配置
物理备份：
开启
关闭
group_1531727761_858077 / 计平-上海集群 / 业务（微信渠道订单(tinasu)） / VHOST（ ）
运营状态：
待运营
运营中
已下架
查询
设置【告警策略】
设置【运营状态】
设置【告警屏蔽】
Host (IP:Port)
使用 链接数
SQL 总数
SQL 完成数
SQL 错误数
(0~5ms) 请求量
(5~20ms) 请求量
(20~30ms) 请求量
(>30ms) 请求量
存活 状态
异常 情况
324 14,860 14,861 0 14,839 22 0 0 正常 正常
328 11,948 11,950 0 11,938 11 0 1 正常 正常
124 13,646 13,646 0 13,588 58 0 0 正常 正常
binlog和冷备的保存天数：
30
保存
总请求量
长事务次数
活跃线程数
活跃连接数
CPU利用率
库数量
替换备机
删除备机

告警



对象	类型	命名	操作
实例	公共	实例告警策略	修改
MySQL	公共	Mysql告警策略	修改
Proxy	公共	Proxy监控策略	修改
Zookeeper	公共	ZK告警策略	修改
多源同步任务	公共	多源同步监控策略	修改
OSS任务	公共	OSS任务流程监控策略	修改
集群容量	公共	容量统计告警	修改
HDFS	公共	HDFS告警策略	修改
KAFKA	公共	Kafka告警策略	修改
Scheduler	公共	Scheduler告警策略	修改
设备资源	公共	设备资源监控策略	修改
OSS服务	公共	OSS服务监控策略	修改
Manager	公共	Manager告警策略	修改

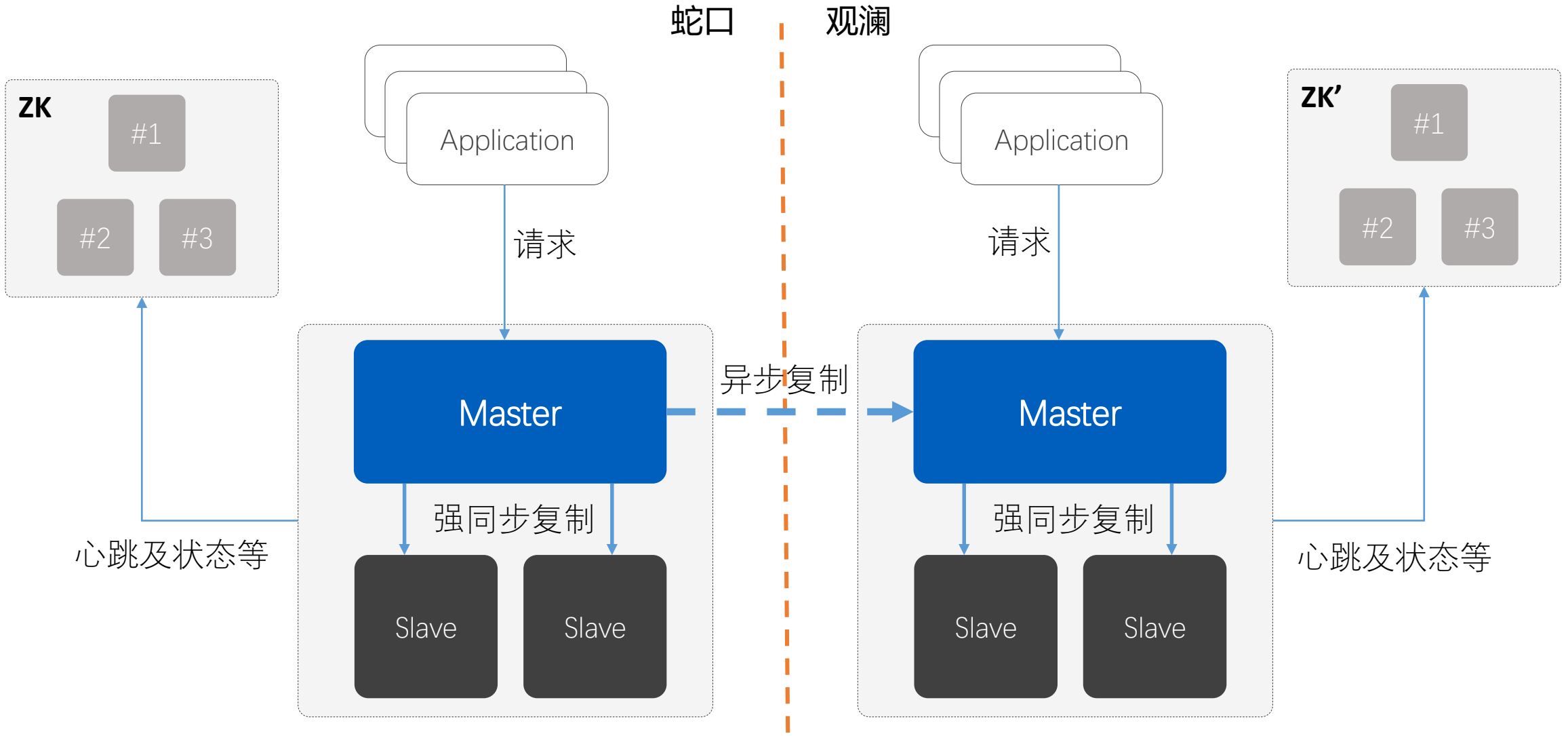
扁鹊-TDSQL数据库诊断



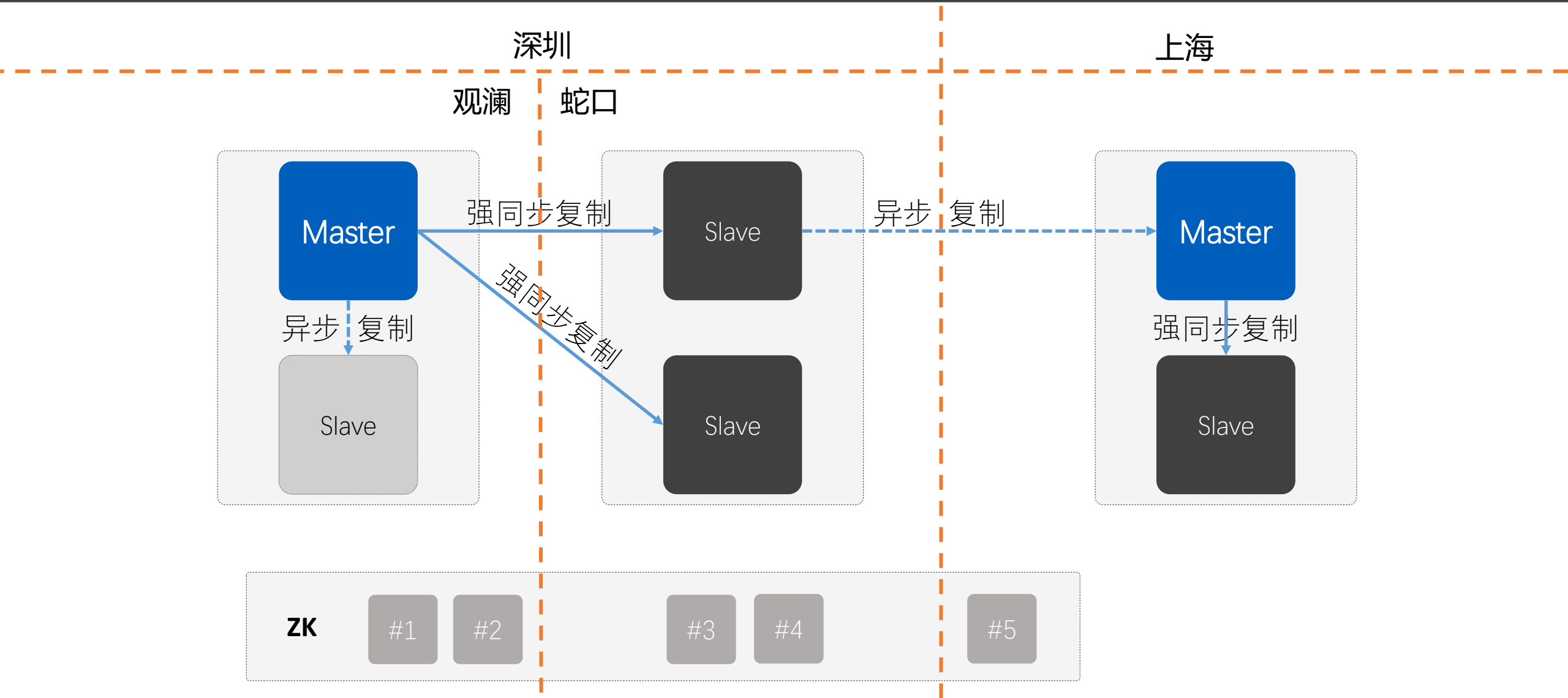
4 部署实践

多地多中心、强同步异步灵活部署

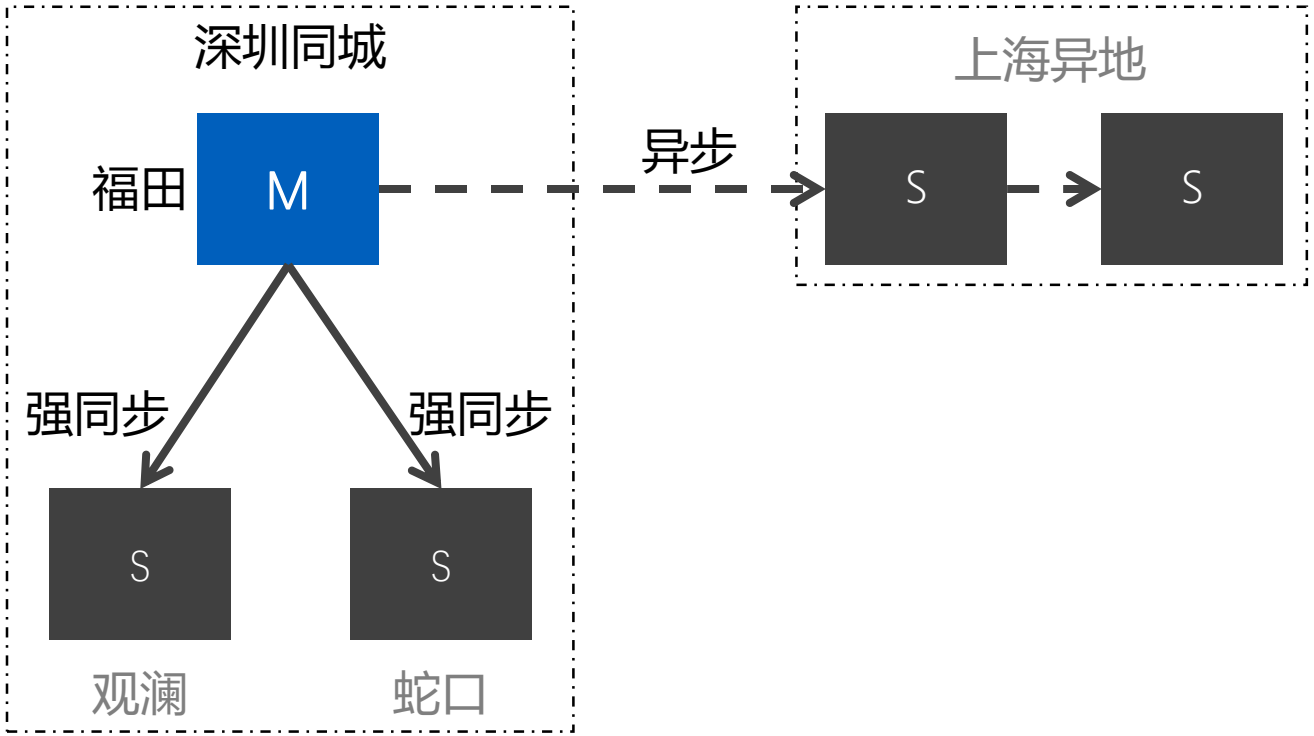
同城主从双中心


蛇口
观澜
ZK
#1
#2
#3
Application
请求
Application
请求
ZK'
#1
#2
#3
Master
异步复制
Master
强同步复制
强同步复制
心跳及状态等
心跳及状态等
Slave
Slave
Slave
Slave

两地三中心


深圳
上海
观澜
蛇口
Master
强同步复制
Slave
异步 复制
Master
异步 复制
强同步复制
Slave
强同步复制
Slave
Slave
ZK
#1
#2
#3
#4
#5

两地四中心 --(自动化切换的强同步架构)



- 同城三中心集群化部署，简化同步策略，运营简单，数据可用性、一致性高
- 单中心故障不影响数据服务
- 深圳生产集群三中心多活
- 整个城市故障可以人工切换

关于「3306π」社区

围绕**MySQL**核心技术，将互联网行业中最重要的数据化解决方案带到传统行业中

囊括**其他开源技术，**redis、MongoDB、Hbase、Hadoop、ElasticSearch、Storm、Spark等

在全面互联网化的大趋势下，将互联网新鲜的核心技术理念带到传统行业里，构建良好交流互动环境

分享干货知识，即便是赞助商，也要求如此，拒绝放水

「3306π」社区，欢迎您的加入

社区公众号

社区QQ群

Database for your business