饿了么异地双活数据库实战

虢国飞

目录

多活难点-同城Or异地

	同城多活	异地多活
整体投入	高（机房投入 + 同城专线）	很高（机房投入 + 异地专线）
实现复杂度	低（依赖垮机房调用）	高（需要减少机房间的交互，清理调用边界）
可以扩展到多机房	中（只能在同城增加机房）	高（可以在全国选择机房，甚至扩展到全球）
服务可用性	低（降低现有可用性）	高（可以应对机房级故障）
对现有架构的影响	低（跨机房调用）	高（业务需要改造）
对服务质量的影响	降低实时性，增加延迟的风险	能够保证实时和服务质量 ✓

多活难点-数据问题

- 错乱
- 冲突
- 环路
- 一致性

多活难点

- 如何做好分流和控制？
- 如何解决跨机房延时（访问&数据）？
- 如何解决数据安全性？

设计原则-业务

1 业务内聚

一个订单的旅单过程在一个机房中完成以减少可能的延迟

2 可用性优先

优先保证系统可用，让用户可以下单吃饭，容忍暂时数据不一致，事后修复

4 业务可感

业务团队修改逻辑，能够识别出业务单元的边界，只处理本单元的数据，打造强大的业务状态机，能发现和纠正错误

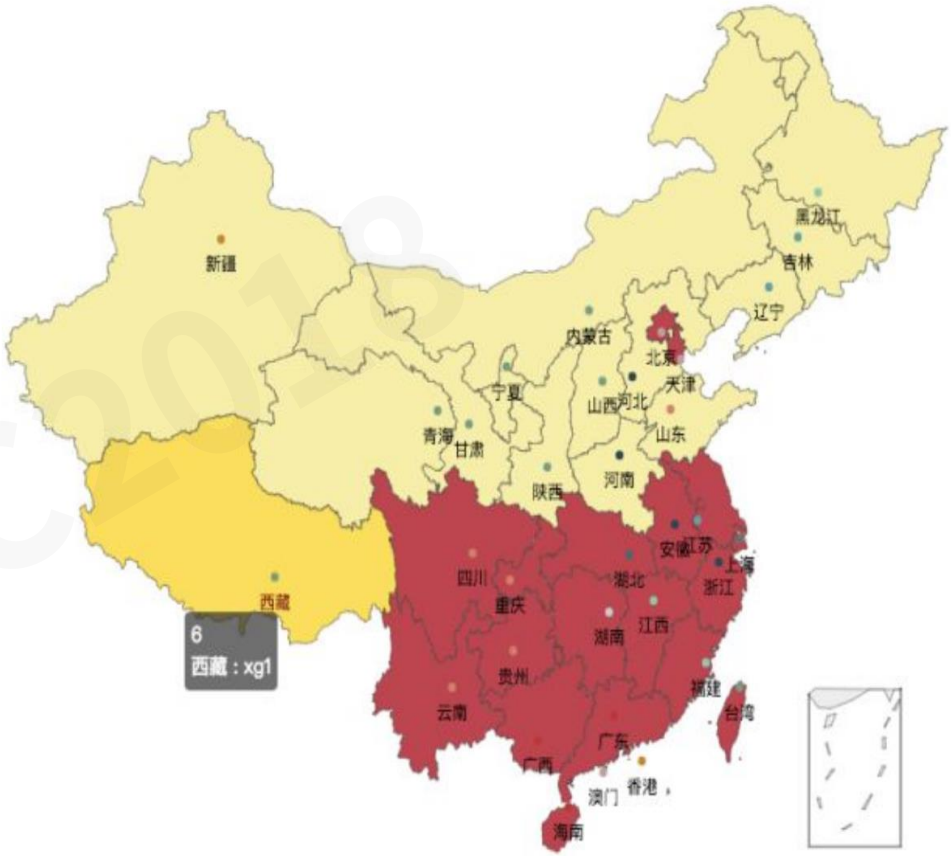
3 保证正确性

在确保可用的情况下，需要对数据做保护以避免错误

设计原则-分区

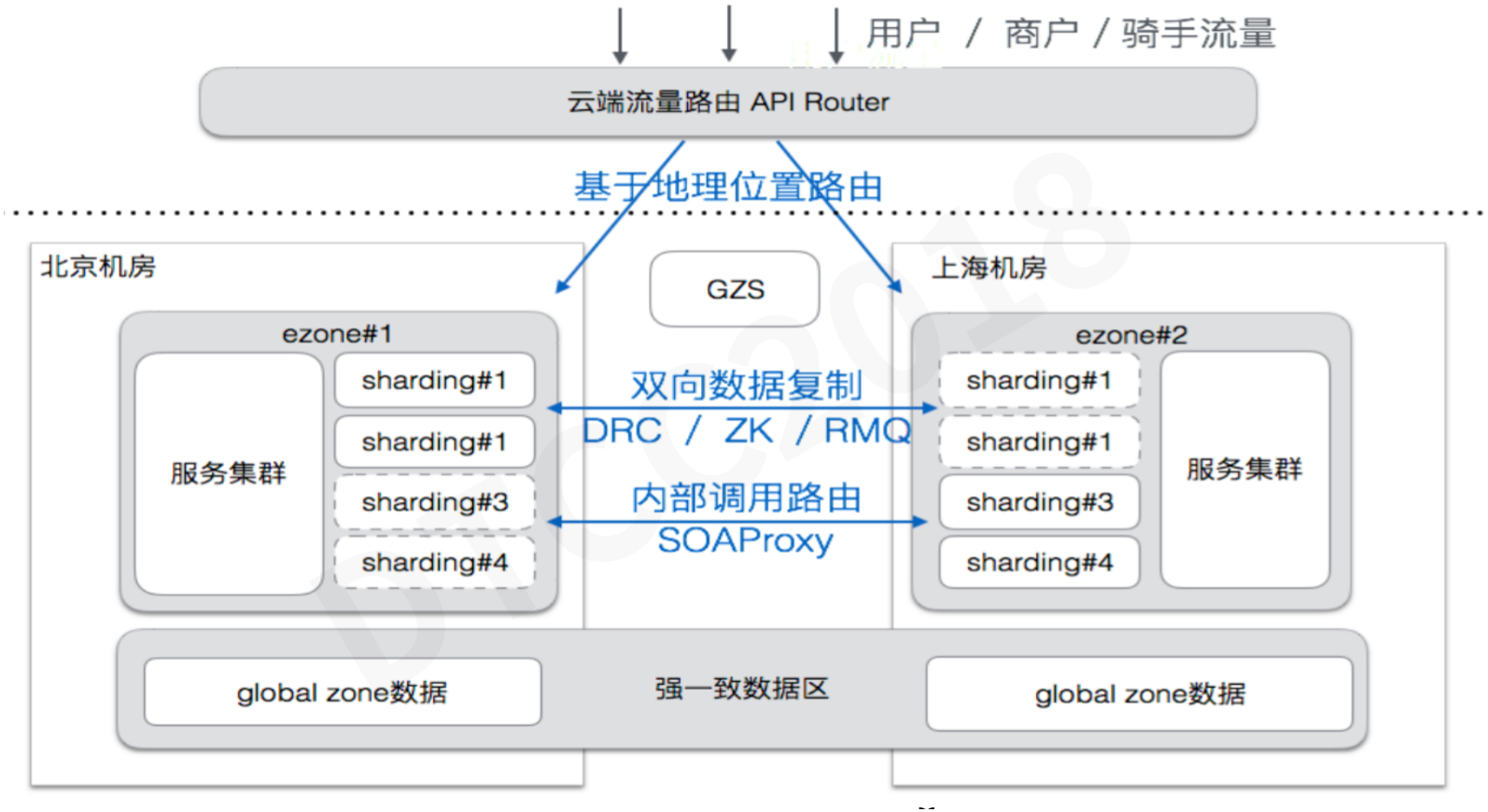
- 用户、商户、骑手（POI）

ezone	sharding ID	商户ID
alta1	1	999999991
	2	999999992
	3	999999993
altb1	4	999999994
	5	999999995

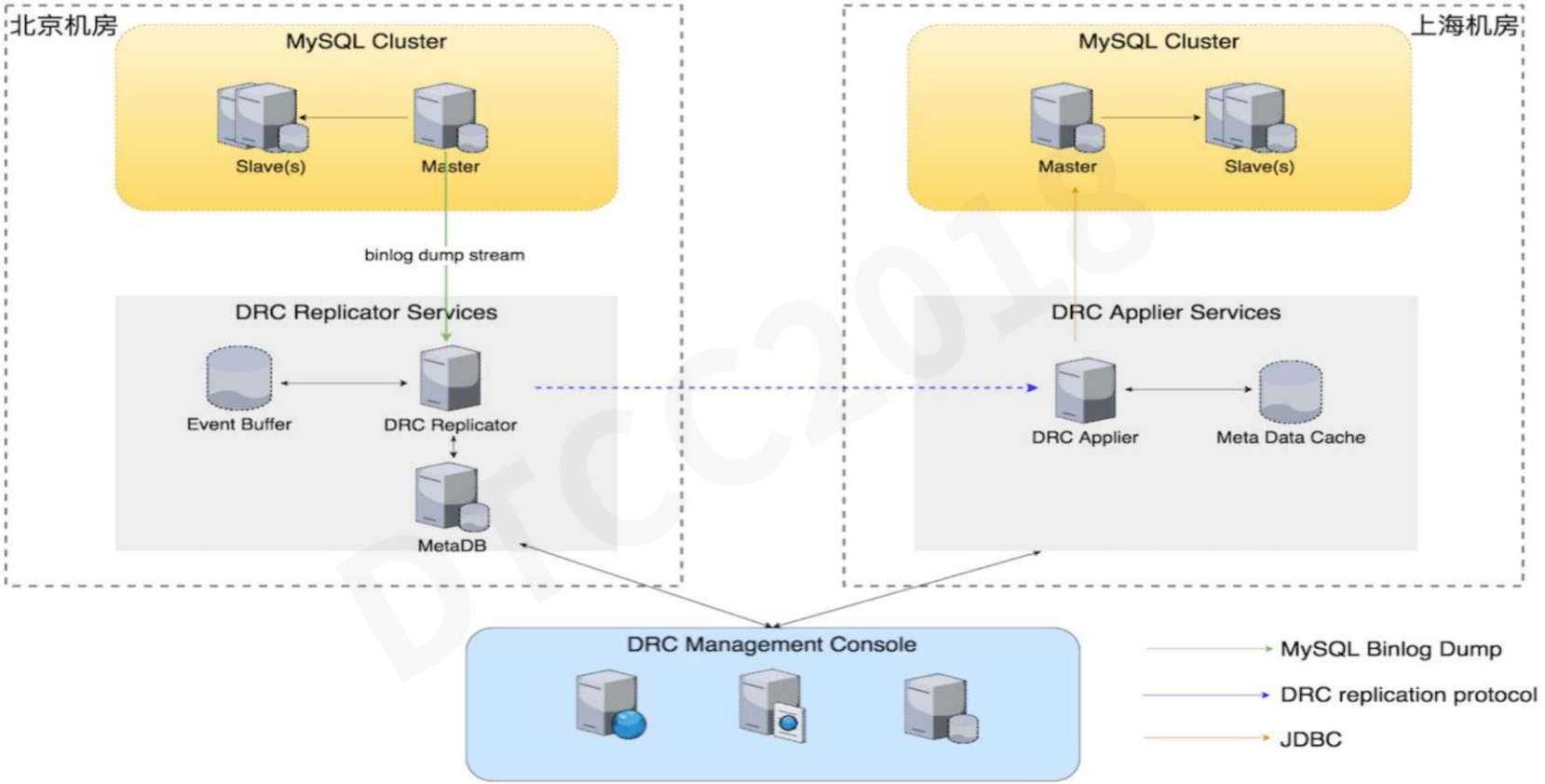


目录

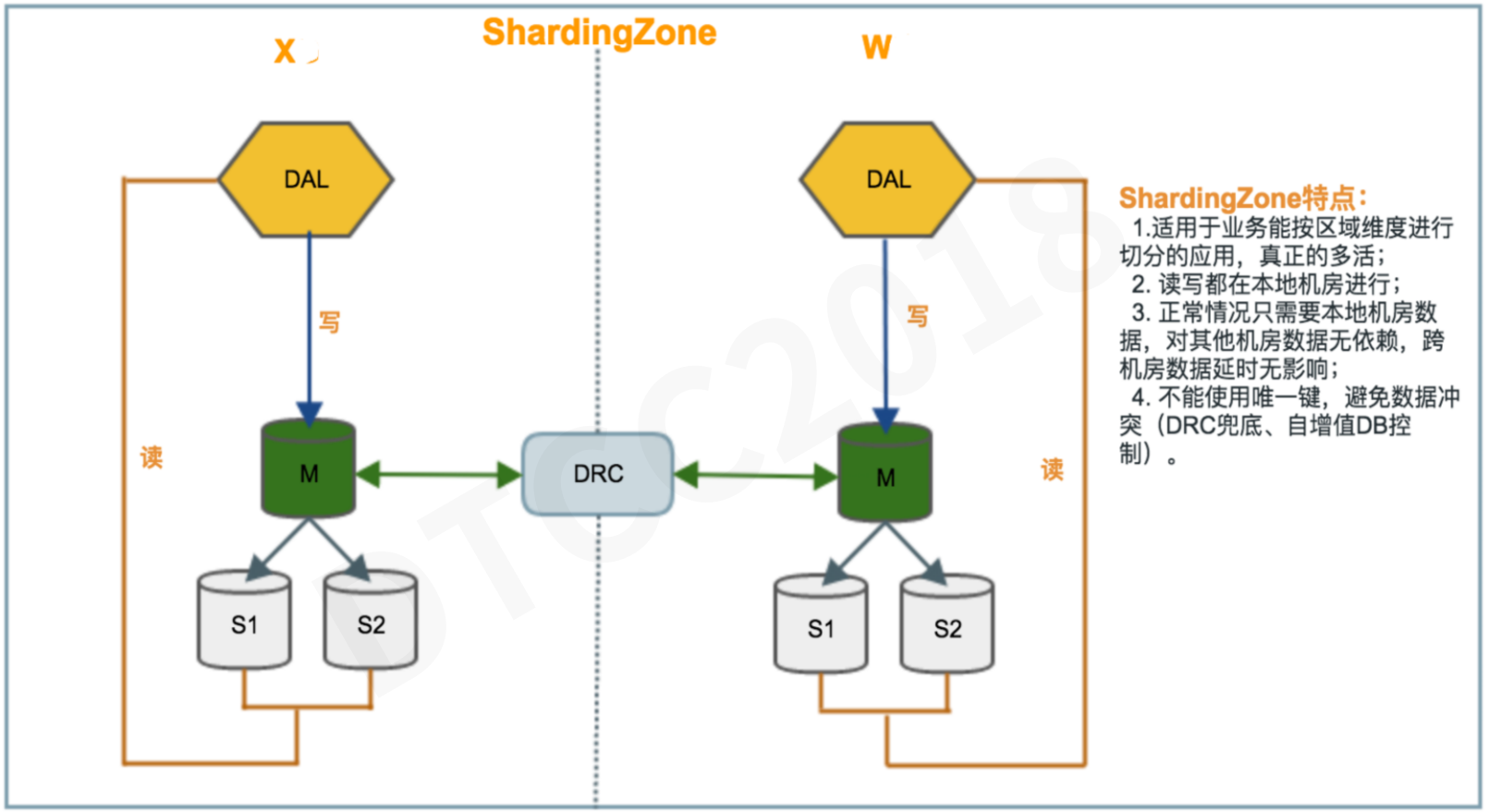
多活架构-Overview



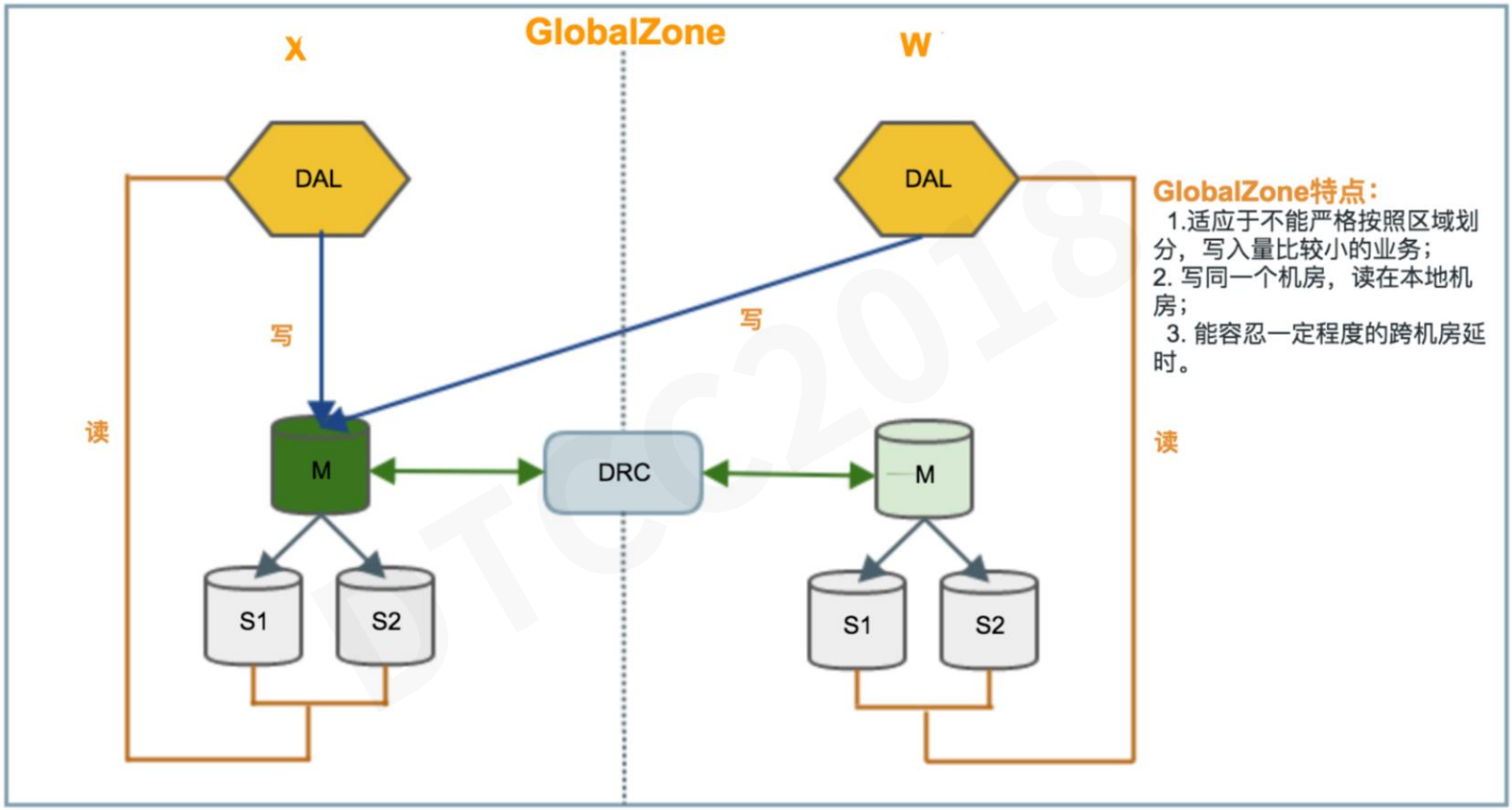
多活架构-DRC



多活架构-DB（SZ）



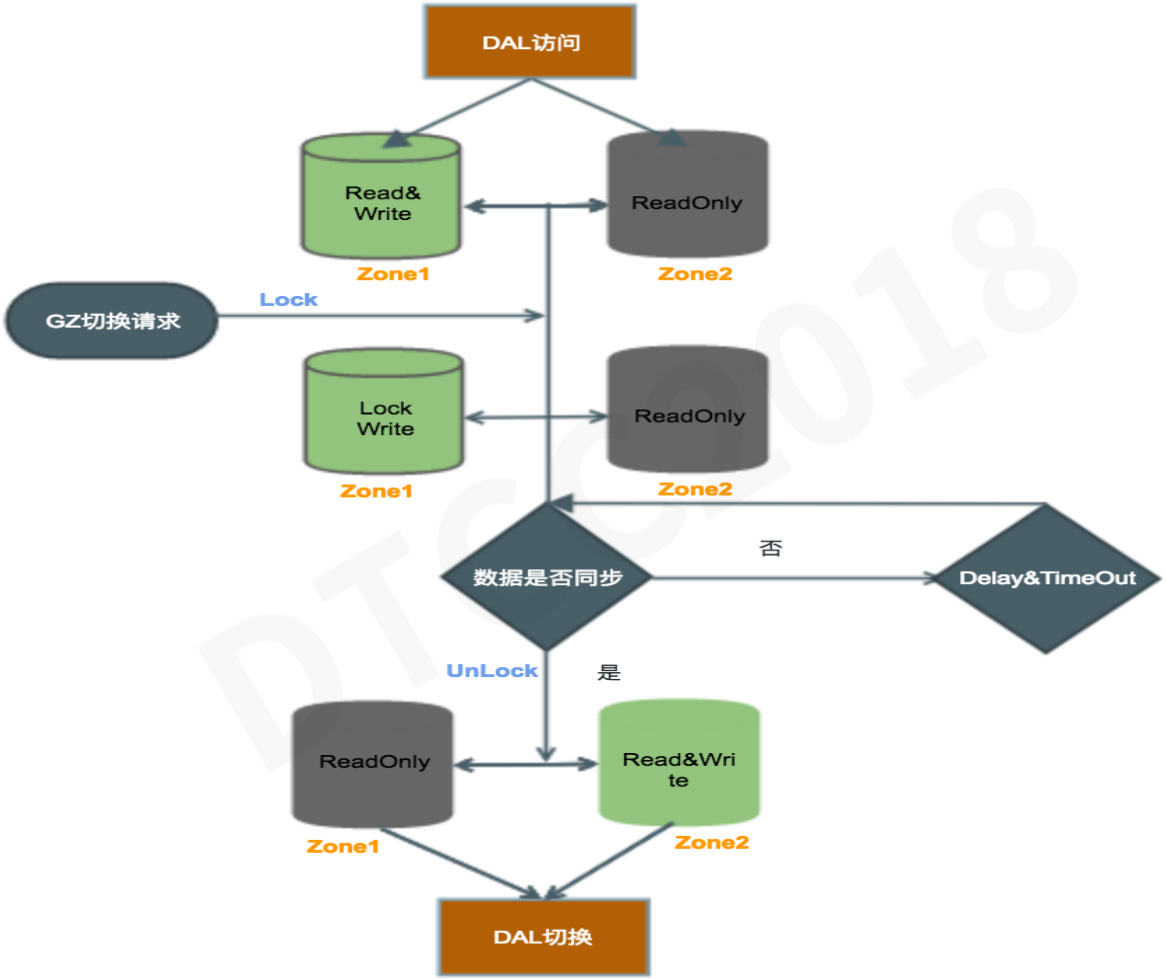
多活架构-DB（GZ）



多活切换-流量

		初始状态 shard:1 ezone: ezone1 status: ACTIVE	BLOCK状态 shard:1 ezone: ezone1 status: BLOCK	RESHARD 状态 shard:1 ezone: ezone2 status: RESHARD	结束状态 shard:1 ezone: ezone2 status: ACTIVE
	API Router	转发	拒绝	转发	转发
ezone1	SOA Proxy	通过	通过	转发	转发
	DAL	通过	拒绝	拒绝	拒绝
ezone2	SOA Proxy	转发	转发	通过	通过
	DAL	拒绝	拒绝	允许下单。拒绝对老订单的 修改	通过

多活切换-DB（GZ）



目录

数据库改造-项目

项目	改造原因	数据量	周期
全量数据导入	测试环境、生产环境数据全量同步	几百TB	两周
表增加DRC字段	表增加毫秒级别的时间戳，方便判断数据有效性	十几万表	五周
PK int改bigint	自增调整防止溢出	十几万表	
FK改bigint	Pk调整防止溢出	几万表	
业务分类迁移	不同类型业务要求放入不同集群	50+ DB	
自增调整	防止自增冲突，每个zone起始值错开	几百套	三周
原生改DRC	原生复制改成DRC复制，支持多写	几百套	
账号网段调整	原来账号限制在一个机房，现在需要支持多个机房	数千账号	
全量参数一致性	各个集群参数必须一致	几百套	
HA全量部署改造	按集群类型调整HA配置	几百套	

数据库改造-比较

项目	改造前	改造后
实例	1200+	2000+
集群	400+	800+
Proxy	800+	1600+
HA	400+	800+
数据量	几百TB	翻倍
DDL	3位数/周	翻倍
DML	2位数/周	不变
机器故障	0.5台/周	2台/周
DBA	?	+1

DBA挑战-数据

- DAL-Reject
- DRC-冲突
- DCP-校验

DBA挑战-数据（DCP）

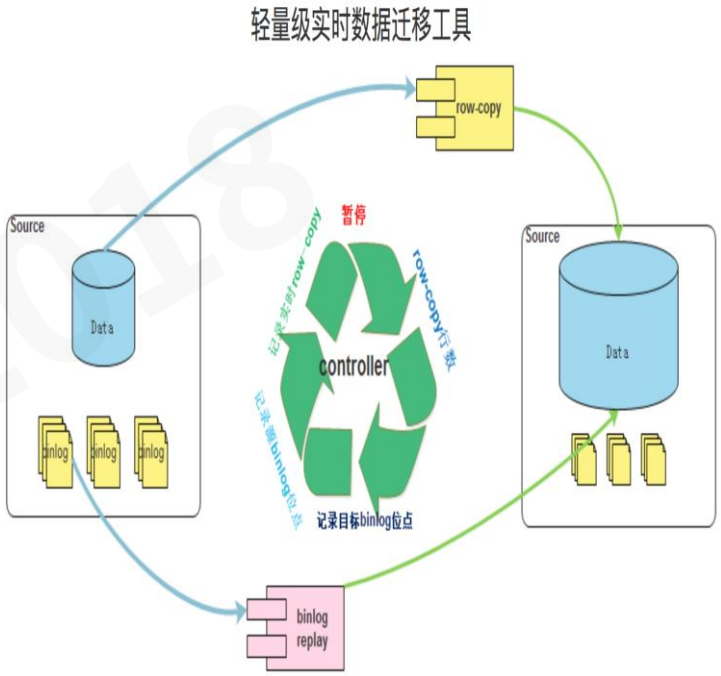
- 变动无需人工干预
- 全量、增量、延时校验、手动校验
- 数据、结构、多维校验
- 延迟、并发、时长
- 修复SQL、配套工具

DBA挑战-数据（DCP)

- 每日几百套集群数据校验
- 日均校验数据60亿+
- 分钟级别校验频率
- 发现和修复数据一致性问题50+

DBA挑战-迁移（D-Bus）

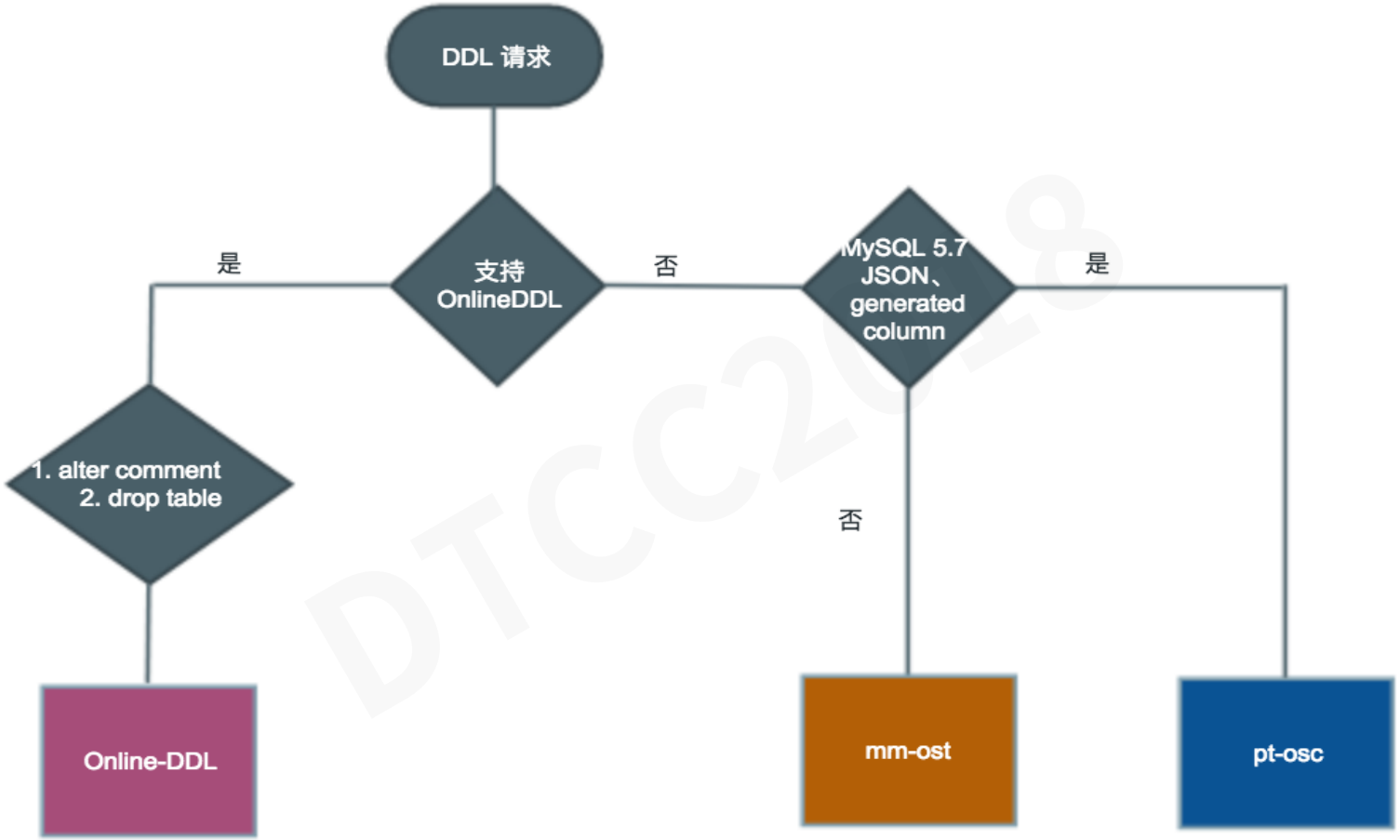
- DB&Table迁移
- 增量、实时同步
- 暂停、断点续传
- 单表、Sharding表数据互转
- 数据校验



DBA挑战-HA

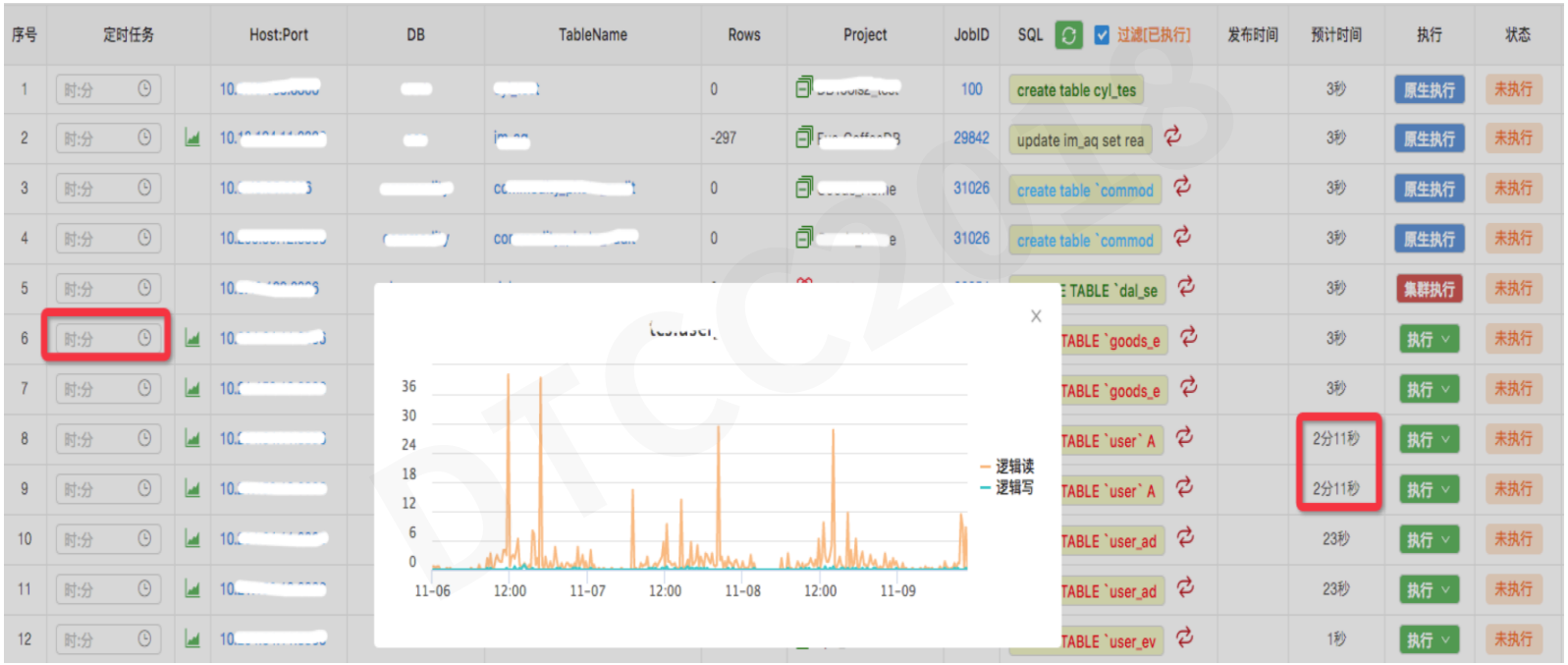
- EMHA 配置、切换、联动（DAL、DRC）

DBA挑战-DDL-工具



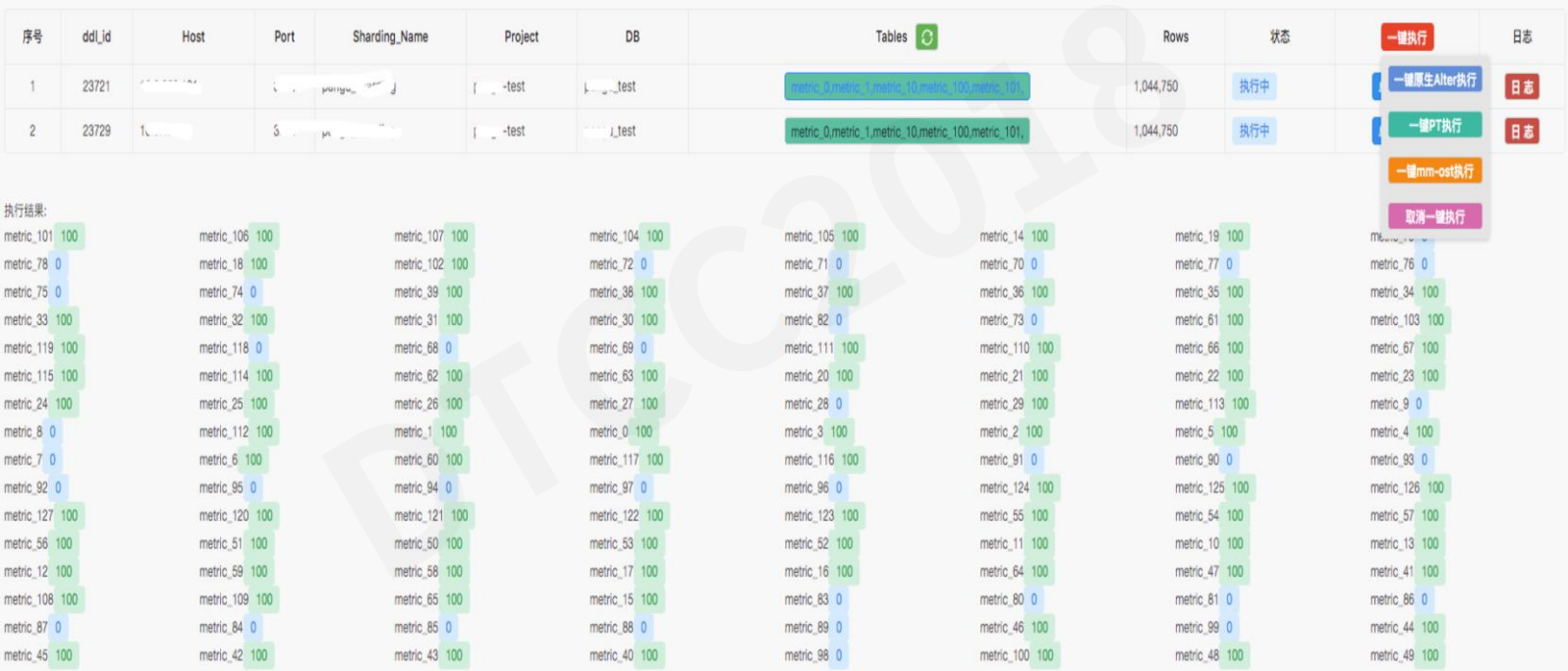
DBA挑战-DDL-控制

- 控制：空间&延时&锁&定时&低峰&风险&时长



DBA挑战-DDL-Sharding

- 类型：多活、非多活 、GlobalZone、多推、Sharding



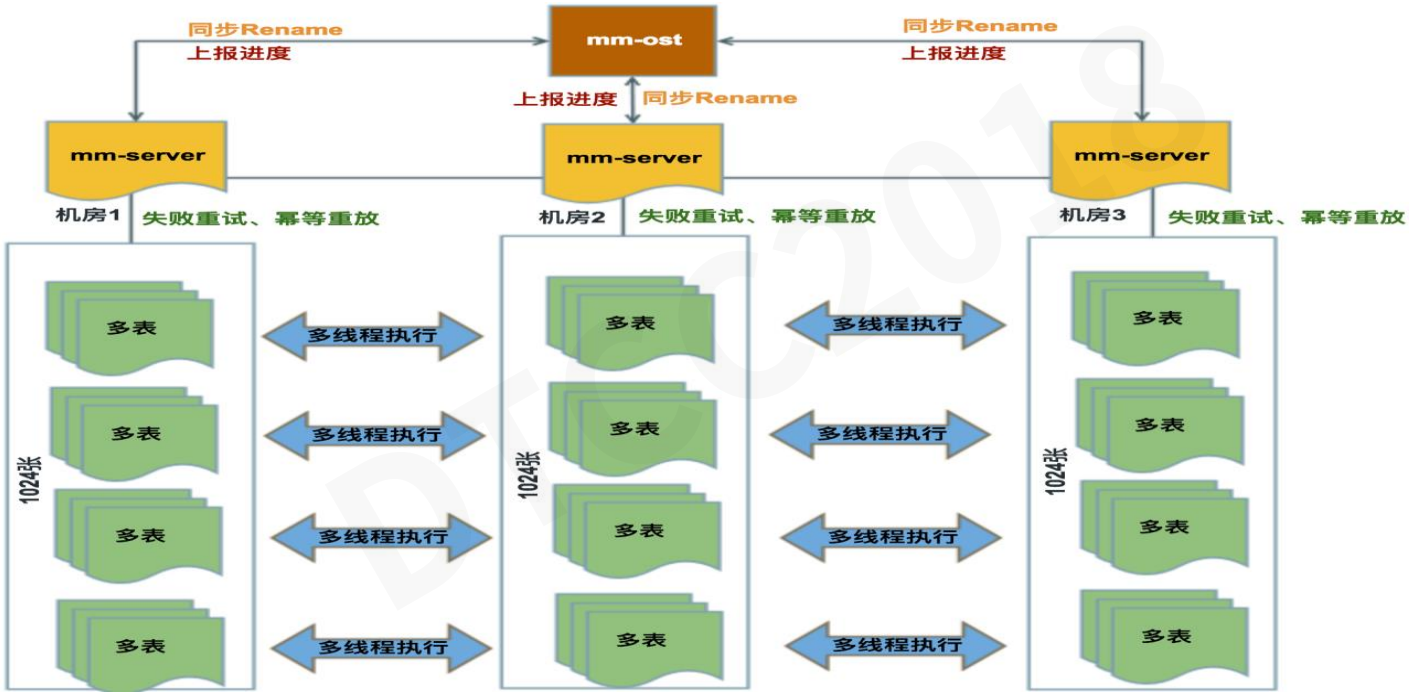
序号	ddl_id	Host	Port	Sharding_Name	Project	DB	Tables	Rows	状态	一键执行	日志
1	23721				-test	_test	metric_0,metric_1,metric_10,metric_100,metric_101,	1,044,750	执行中	一键原生Alter执行	日志
2	23729				-test	_test	metric_0,metric_1,metric_10,metric_100,metric_101,	1,044,750	执行中	一键PT执行	日志
										一键mm-ost执行	
										取消一键执行	

执行结果:

metric_101 100	metric_106 100	metric_107 100	metric_104 100	metric_105 100	metric_14 100	metric_19 100	
metric_78 0	metric_18 100	metric_102 100	metric_72 0	metric_71 0	metric_70 0	metric_77 0	metric_76 0
metric_75 0	metric_74 0	metric_39 100	metric_38 100	metric_37 100	metric_36 100	metric_35 100	metric_34 100
metric_33 100	metric_32 100	metric_31 100	metric_30 100	metric_82 0	metric_73 0	metric_61 100	metric_103 100
metric_119 100	metric_118 0	metric_68 0	metric_69 0	metric_111 100	metric_110 100	metric_66 100	metric_67 100
metric_115 100	metric_114 100	metric_62 100	metric_63 100	metric_20 100	metric_21 100	metric_22 100	metric_23 100
metric_24 100	metric_25 100	metric_26 100	metric_27 100	metric_28 0	metric_29 100	metric_113 100	metric_9 0
metric_8 0	metric_112 100	metric_1 100	metric_0 100	metric_3 100	metric_2 100	metric_5 100	metric_4 100
metric_7 0	metric_6 100	metric_60 100	metric_117 100	metric_116 100	metric_91 0	metric_90 0	metric_93 0
metric_92 0	metric_95 0	metric_94 0	metric_97 0	metric_96 0	metric_124 100	metric_125 100	metric_126 100
metric_127 100	metric_120 100	metric_121 100	metric_122 100	metric_123 100	metric_55 100	metric_54 100	metric_57 100
metric_56 100	metric_51 100	metric_50 100	metric_53 100	metric_52 100	metric_11 100	metric_10 100	metric_13 100
metric_12 100	metric_59 100	metric_58 100	metric_17 100	metric_16 100	metric_64 100	metric_47 100	metric_41 100
metric_108 100	metric_109 100	metric_65 100	metric_15 100	metric_83 0	metric_80 0	metric_81 0	metric_86 0
metric_87 0	metric_84 0	metric_85 0	metric_88 0	metric_89 0	metric_46 100	metric_99 0	metric_44 100
metric_45 100	metric_42 100	metric_43 100	metric_40 100	metric_98 0	metric_100 100	metric_48 100	metric_49 100

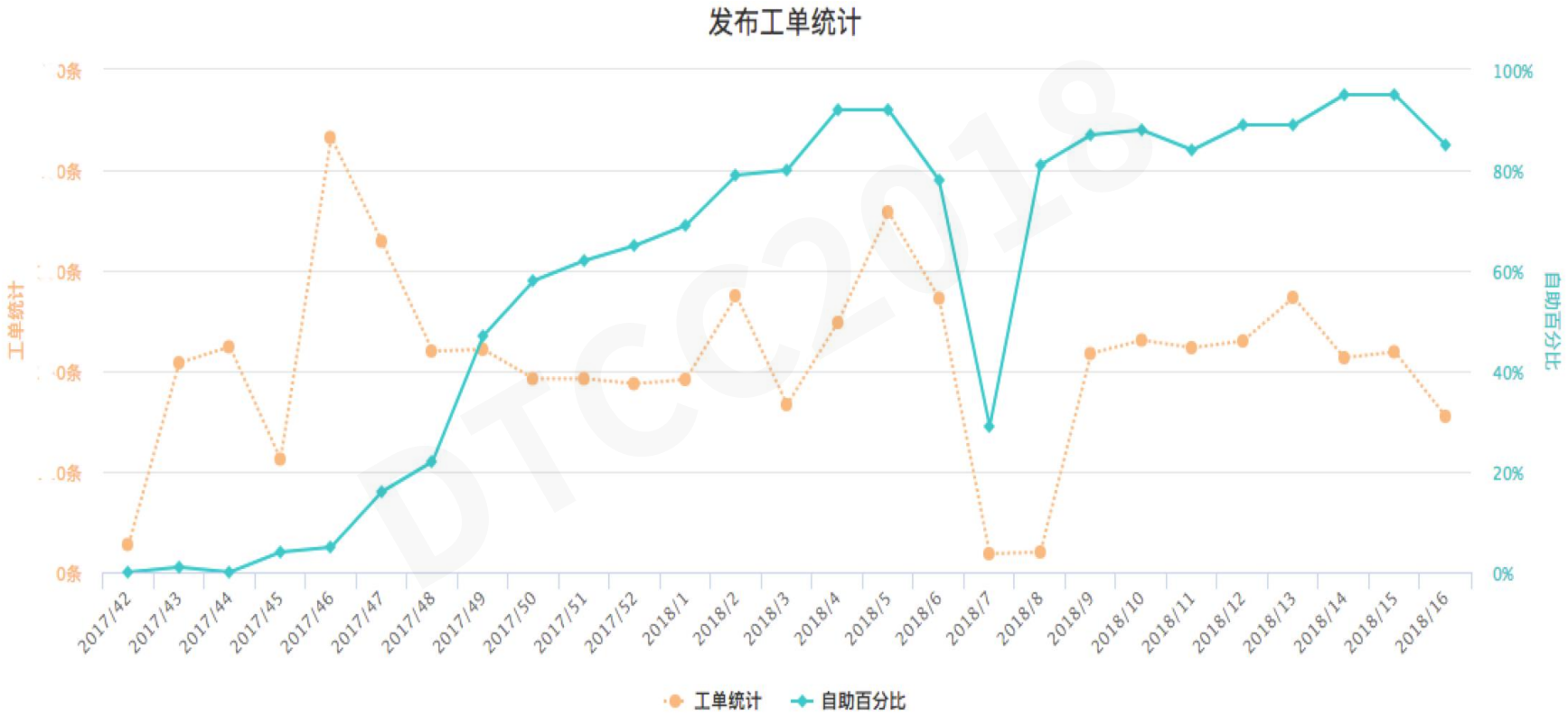
DBA挑战-DDL-多机房

- Mm-ost： 多机房一致&延时控制3-5s



DBA挑战-DDL-自助

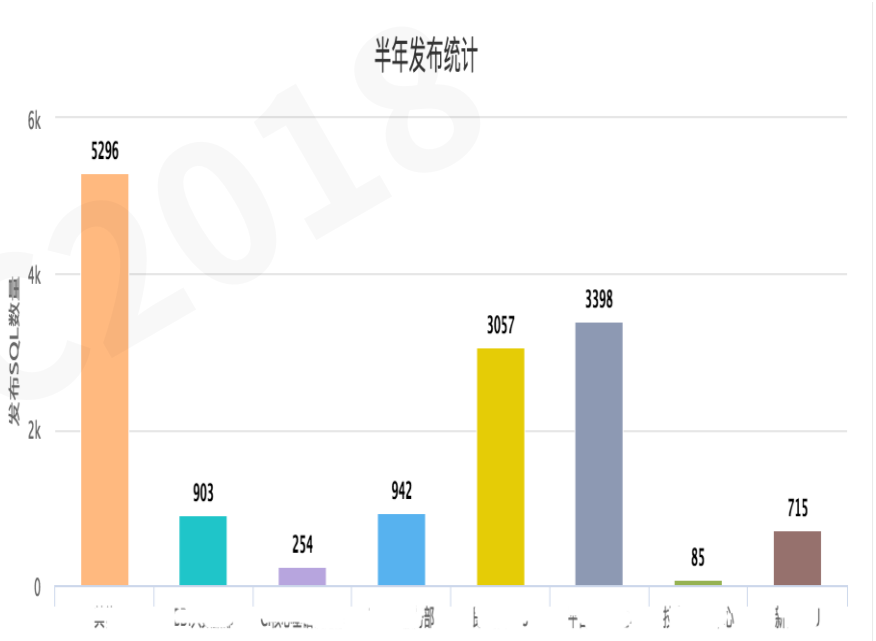
- 研发自助：比例超90%



DBA挑战-DDL

数量：

- 多活工单：4位数/周
- DDL表：4~5位数/周
- 自动/人工：8：2
- 研发自助：90%



目录

收益

- 打破单机房（地域）容量瓶颈
- 不受单机房（地域）故障影响
- 动态调整各机房流量
- Online维护（GZS、DAL、DRC、D-Bus、DCP）

展望

- 多个机房
- Data-Sharding
- **自动动态扩缩容**
- 多机房强一致

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168