China · Shenzhen

腾讯海量监控包袱与创新

聂鑫

腾讯社交网络运营部运维负责人

SPEAKER

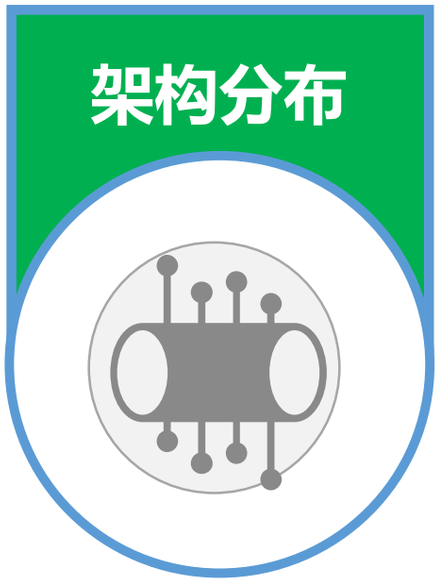
INTRODUCE

聂鑫

- 从开发到运维，伴随腾讯社交网络运营部成长的十年，负责过腾讯社交产品所有业务运维工作，目前主要负责QQ、空间等产品运维团队管理工作。

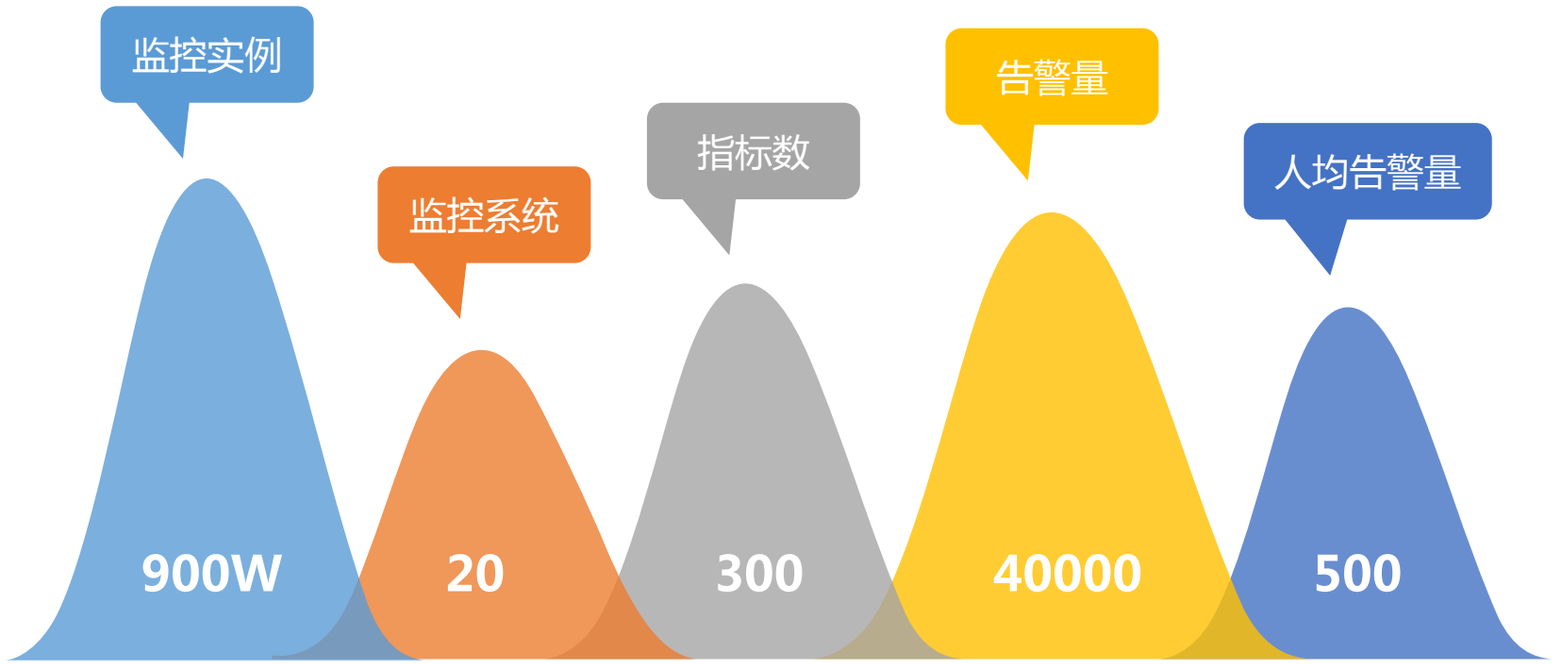
- 经历多个业务产品的诞生到蓬勃，伴随着运维团队的成长和成熟，见证着腾讯一代代运营技术的创新和发展。作为运维界老兵有好多故事想和大家讲，也特别愿意听听各位经历的酸甜苦辣。

腾讯 技术运营重点



短信告警 **5** 万条/天

单人最高1500条/天



- 腾讯正在做哪些监控
- 有哪些不一样的地方
- 有哪些值得关注点

正在做哪些监控

腾讯织云
CLOUD
MANAGEMENT

在监控领域有三个主题{ 快、准、全 }
他们永远是矛盾的，调和矛盾成了运维
技术或艺术

监控体系演进

2006
N:网络质量监控
B:网管基础监控
A:自动化测试
M:模块间调用
S:测速系统

2007
W:站点分析系统
m:模块监控
L:容量管理

2009
I:L5组件监控
Y:一致性Monitor特性监控

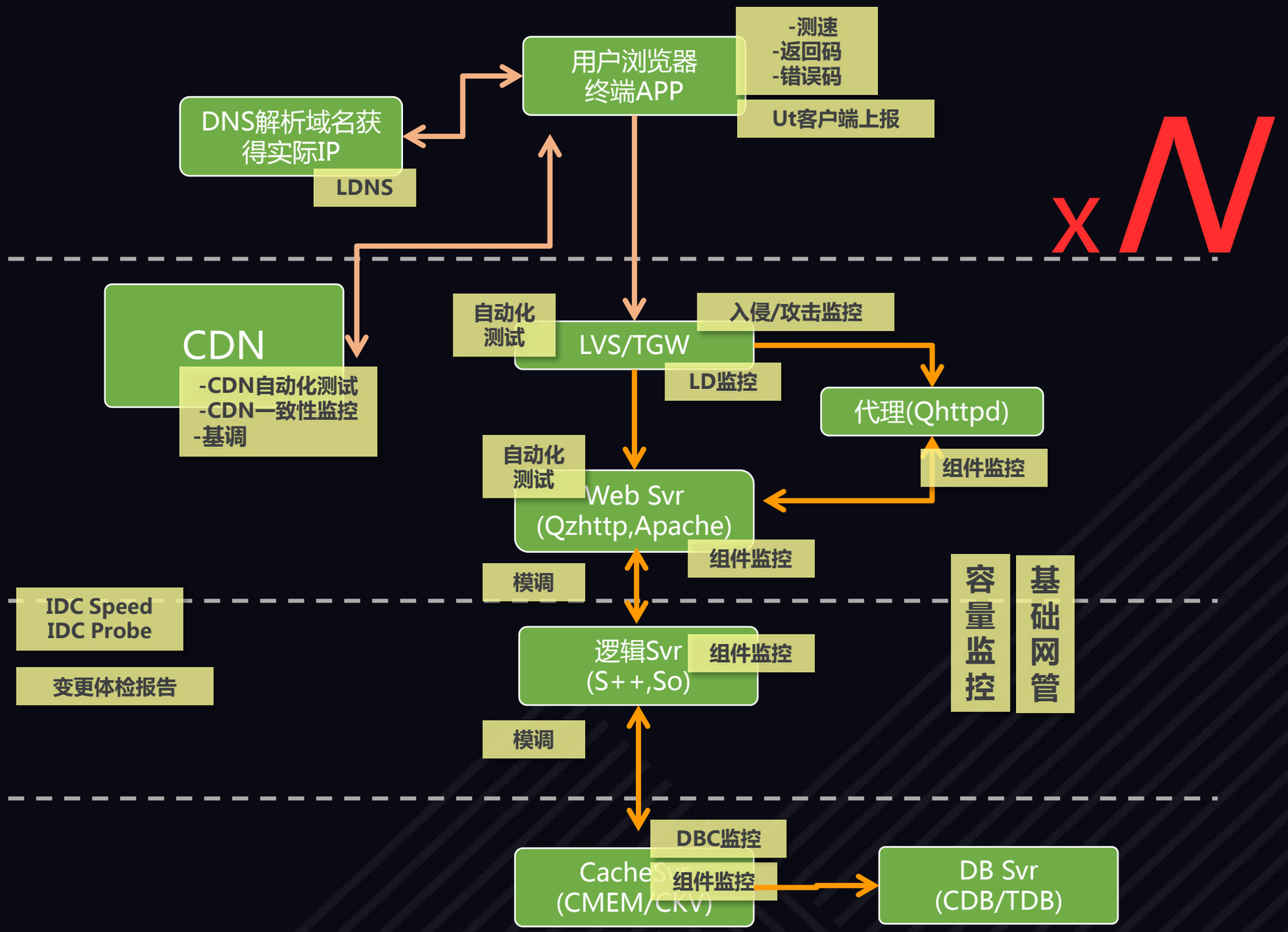
2011
S:QZ组件监控
F:设备特性监控
R:返回码监控

2013
C:CDN监控
P:Ptlogin监控
D:存储质量
H:客户端环境
R:ROOT根源分析

2014
I:华佗移动端分析
Q:舆情监控
G:日志大数据分析
U:UIN染色分析

2015
DLP核心指标
Habo多维分析

2016
Q调拨测监控
移动端卡慢
全链路日志



覆盖完整

用户端监控

- 测速
- 返回码
- 自动化测试
- 基调
- 移动分析(mua)
- html5

业务侧监控

- 核心产品指标
- 各纬度业务指标
- 攻击防御
- 舆情监控

服务内监控

- 模块间调用
- L5失败率
- 组件监控
- 强制一致性
- Gslb、lvs

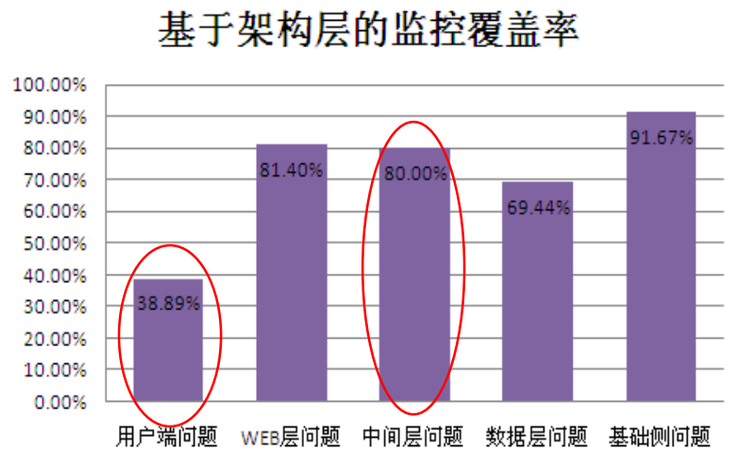
基础资源

- 丢包断线
- 死机重启
- 硬件故障
- 容量监控

多、大、杂

业务增长 VS 监控系统发展

	2009年	2010年	2014年	2017年
主要监控系统数	9	11	**20+**	18
主要监控指标数	132	178	**300+**	400
监控实例数	-	45w	**900w**	2000w
基础告警数/天	192	300	3000+	5000+
业务告警数/天	362	390	**3.9w**	4.97w
个人告警量/天	-	**最大:177**	**500+**	**1500**
(包含运维开发)		**平均:13**		**184**



- 客户端、数据层监控覆盖率弱;
- 系统建设离散，监控数据分散;
- 综合分析能力弱，定位时间长;
- 告警数量过多;

有哪些不一样的地方

腾讯织云
CLOUD
MANAGEMENT

放下包袱来 **创新**

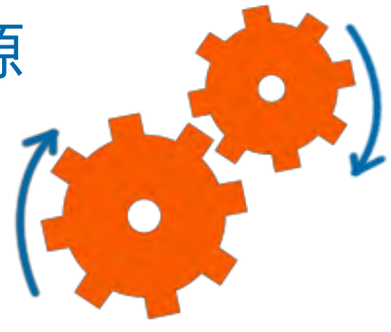
不是破旧立新而是尊重历史

坚决优化历史演进中的架构落后

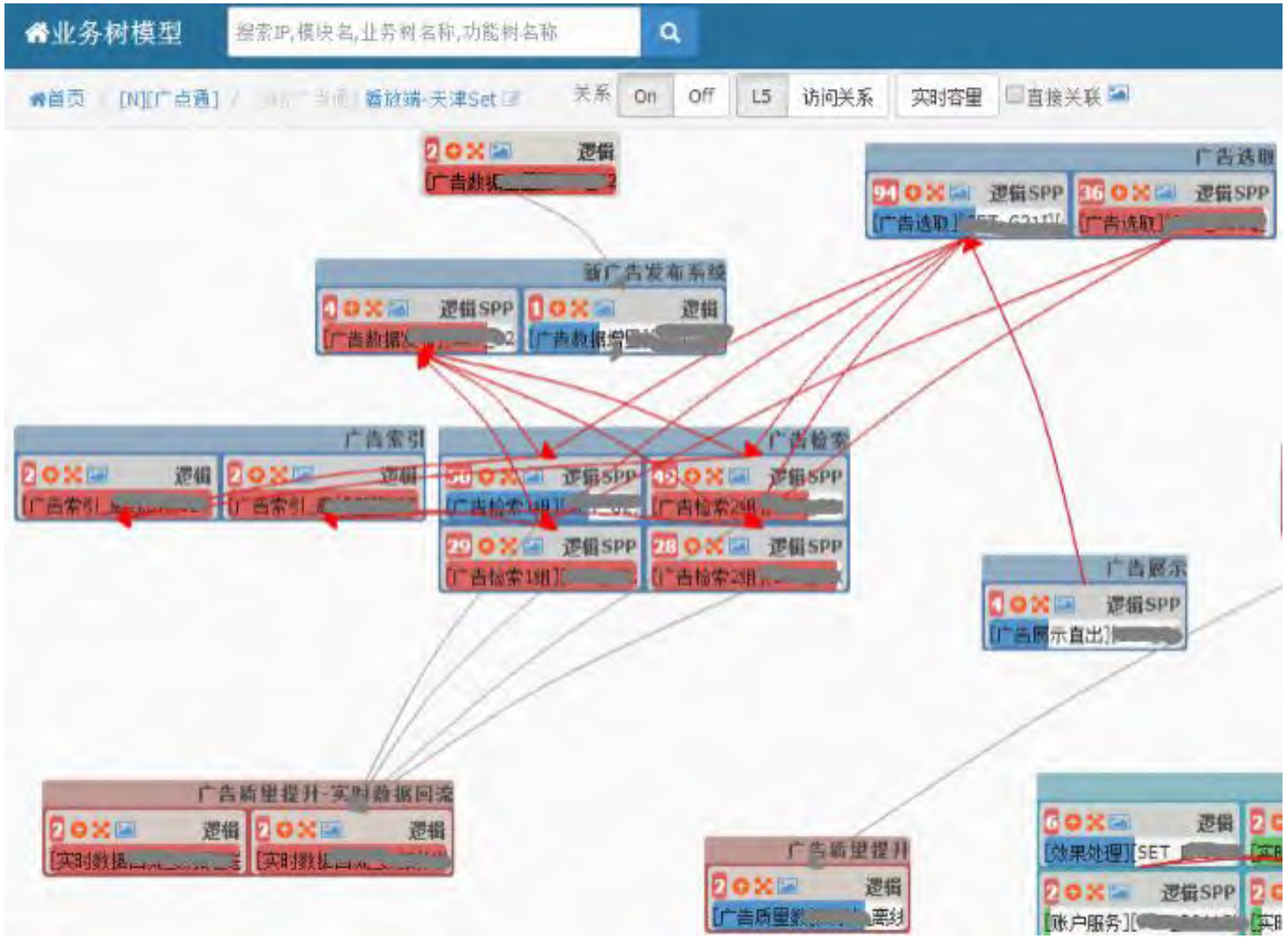
ROOT 根源智能分析法

腾讯织云 CLOUD MANAGEMENT

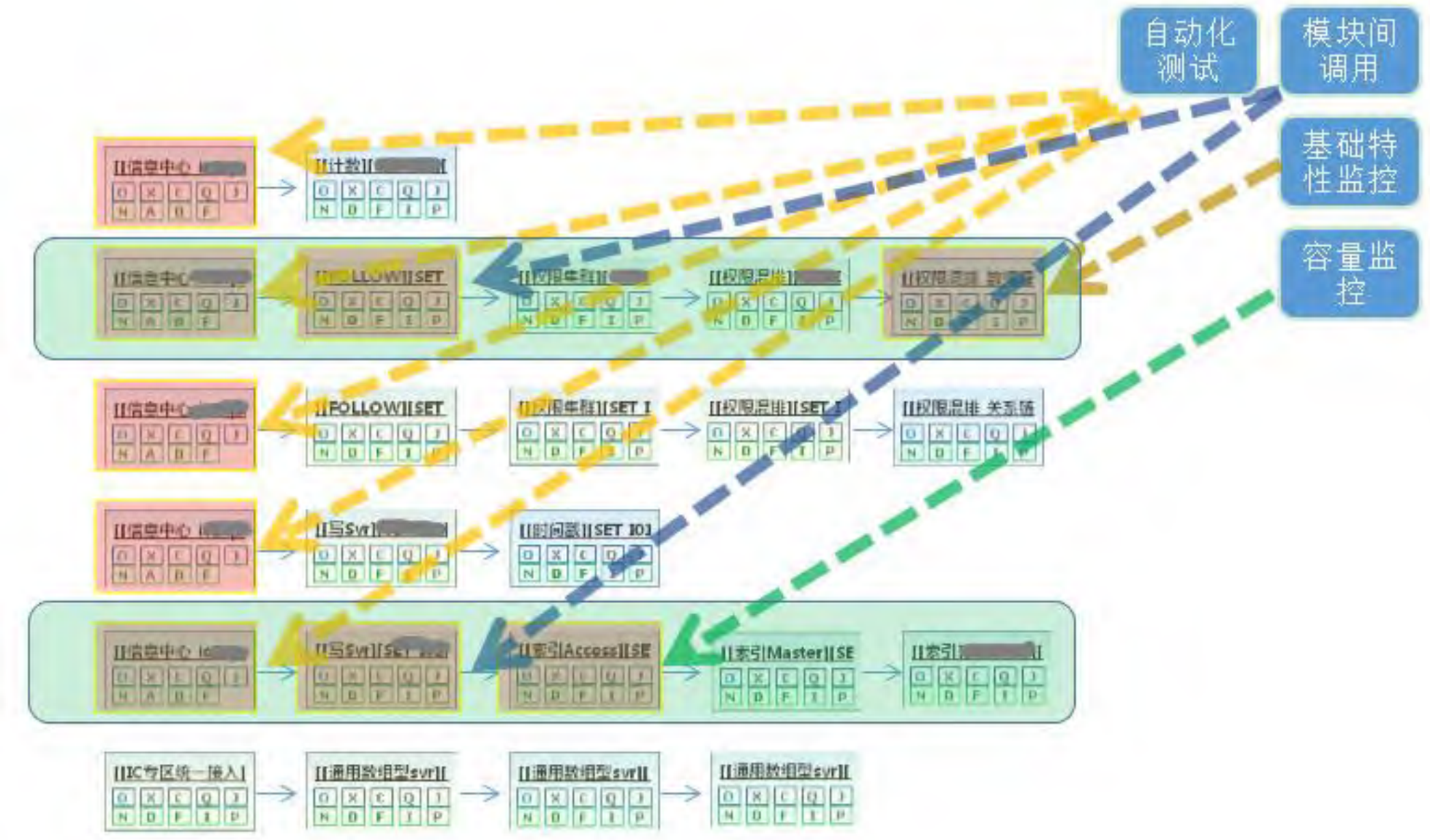
基于**业务架构**，结合**数据流关系**，通过**时间相关性**、**面积权重**等算法，将监控告警进行**筛选分类**，发掘有**业务价值**的告警，并直接分析给出**告警根源**



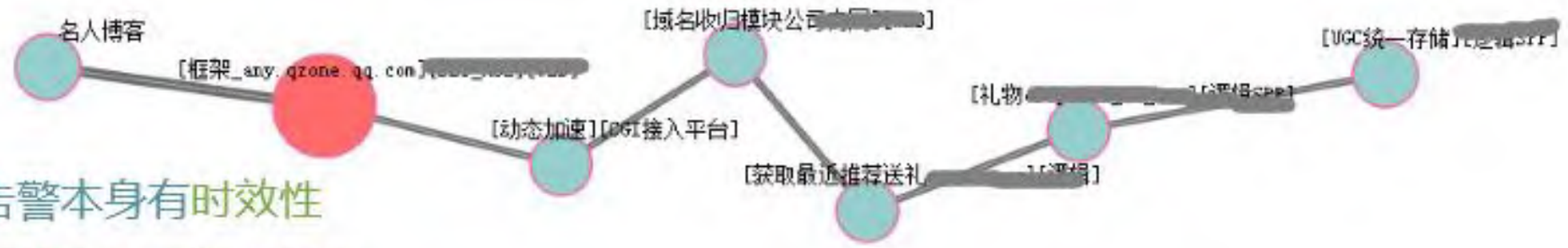
ROOT原理 ： 多维关系降维



ROOT原理 ： 告警叠加


自动化测试
模块间调用
基础特性监控
容量监控
[[FOLLOW]]SET
[[写Svr]]SET_IO
[[索引Access]]SE
[[索引Master]]SE
[[IC专区统一接入]
[[通用数组型svr]]
[[通用数组型svr]]
[[通用数组型svr]]

时间片与时间相关性

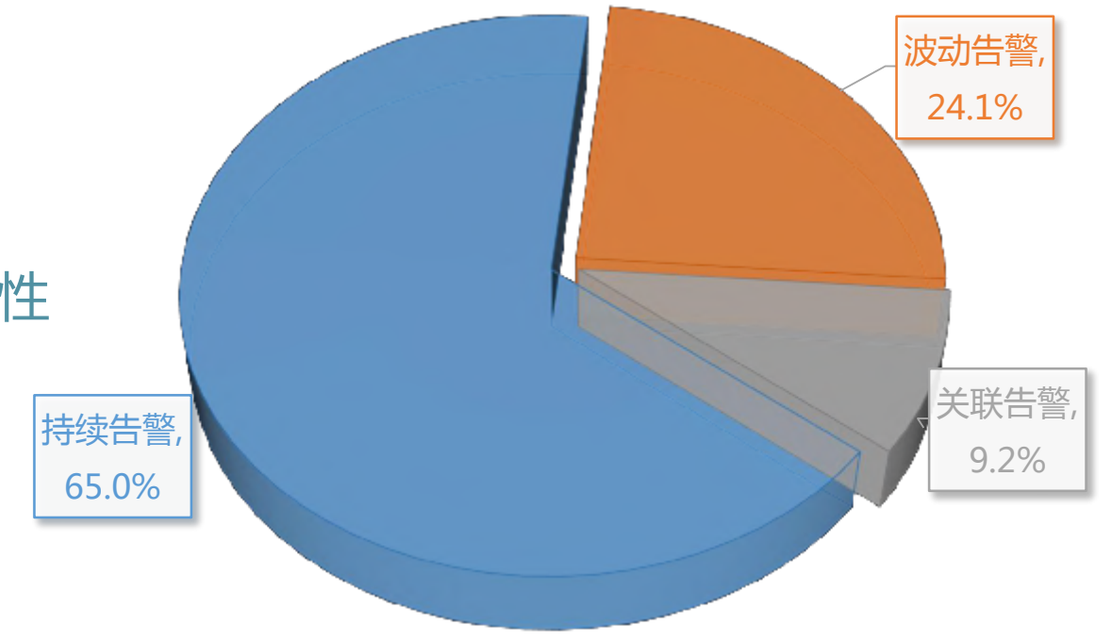


- 告警本身有时效性
- 时效性源于告警延时
- 连续性可能是干扰
- 链路相关性和时间相关性一起决定准确性

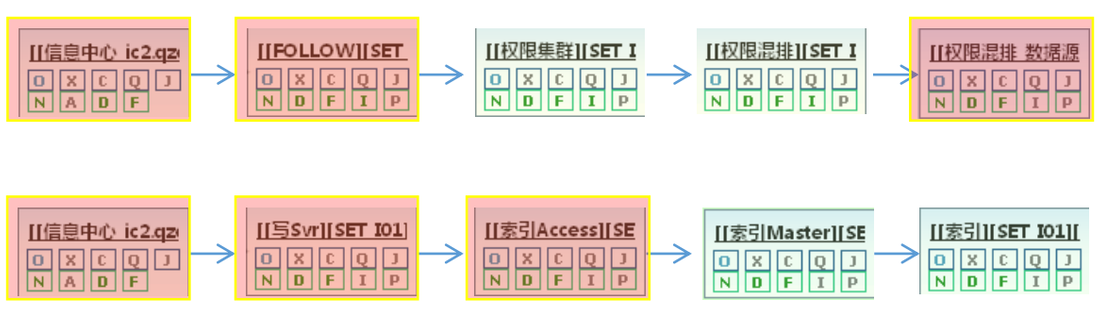
分类筛选和选择性处理

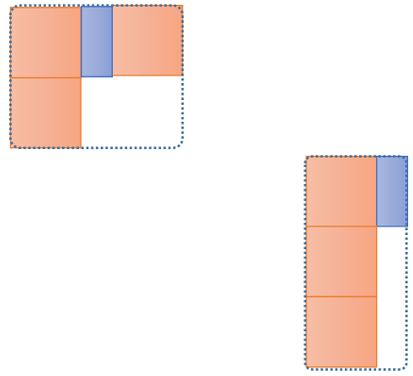
- 原因告警、现象告警
  - 原因告警 ：往往是造成故障的根源，却往往无需处理
  - 现象告警 ：故障的结果，往往看不出根源，需要分析

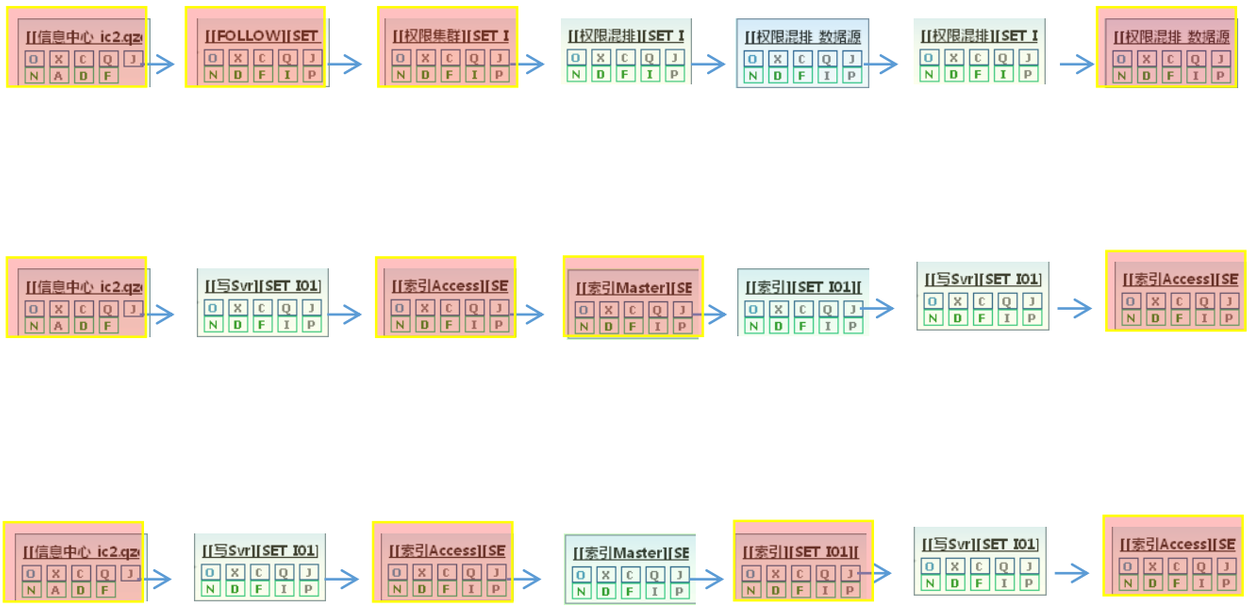
- 持续告警、波动告警、关联告警
  - 持续告警 ：不紧急、不重要
  - 波动告警 ：业务重要性决定告警重要性
  - 关联告警 ：有因有果，即时处理

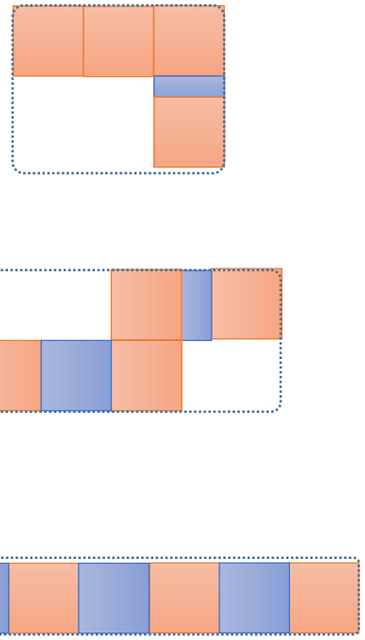


权重与面积算法


[[信息中心 ic2.qz
[[FOLLOW][SET
[[权限集群][SET I
[[权限混排][SET I
[[权限混排 数据源
[[信息中心 ic2.qz
[[写Svr][SET I01]
[[索引Access][SE
[[索引Master][SE
[[索引][SET I01]




[[信息中心 ic2.qz
[[FOLLOW][SET
[[权限集群][SET I
[[权限混排][SET I
[[权限混排 数据源
[[信息中心 ic2.qz
[[写Svr][SET I01]
[[索引Access][SE
[[索引Master][SE
[[索引][SET I01][
[[写Svr][SET I01]
[[索引Access][SE
[[信息中心 ic2.qz
[[写Svr][SET I01]
[[索引Access][SE
[[索引Master][SE
[[索引][SET I01][
[[写Svr][SET I01]
[[索引Access][SE



算法案例

1、链路中告警模块数=1

长=1（只有一个模块告警时固定为1），宽=（1+告警模块所在链路的序号除以链路总模块数），面积=长*宽=1*（1+（iarr+1）/lnkcout）*100

- a、1-0-0-0, 权重面积=1*（1+（0+1）/4）*100=125；
- b、0-1-0-0, 权重面积=1*（1+（1+1）/4）*100=150；
- c、0-0-0-1, 权重面积=1*（1+（3+1）/4）*100=200；

备注：链路中只有一个模块告警，并且结合业务链路生成的特性，告警模块越靠后，权重面积越大；

2、链路中告警模块数>1

长=链路中连着告警模块的最大个数(iarrmax)，宽=连着或不连着告警模块宽都为1+1/(连着不告警的模块个数)，面积=长*宽=iarrmax*(1+1/N+...)*100

- a、1-0-0-0-1, 权重面积=1*（1+1/3+1）*100=233；
- b、1-0-0-1-0, 权重面积=1*（1+1/2+1）*100=250；
- c、1-1-0-0-1, 权重面积=2*（1+1/2+1）*100=500；
- d、1-1-0-1-0, 权重面积=2*（1+1/1+1）*100=600；
- e、1-1-0-0-1-1-0-0-1, 权重面积=2*（1+1/2+1+1/2+1）*100=800；
- f、1-1-0-0-1-1-0-1-1, 权重面积=2*（1+1/2+1+1/1+1）*100=900；
- g、1-1-1-0-1-0-0-1-1, 权重面积=3*（1+1/1+1+1/2+1）*100=1350；

3、特殊情况：

1、链路中，前面模块都没有告警，但最后模块连着告警（相当于链路中全模块告警），权重面积*10；

2、链路中，模块全告警，权重面积*10；

- a、0-0-0-1-1, 权重面积=（2*1*100）*10=2000；
- b、1-1-1-1-1, 权重面积=（5*1*100）*10=5000；

腾讯织云
CLOUD
MANAGEMENT

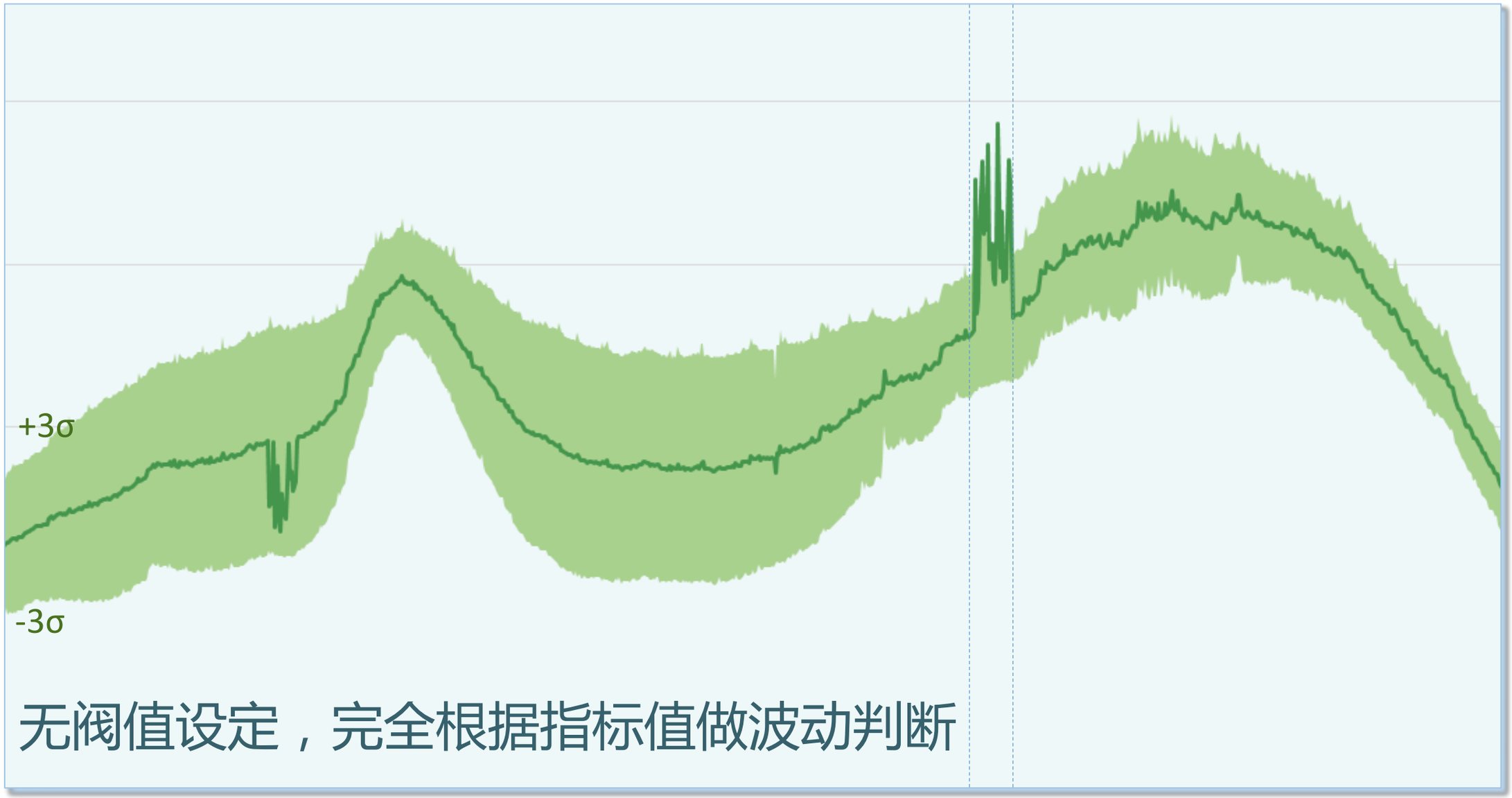
DLP

业务生死指标

衡量业务死、活的指标

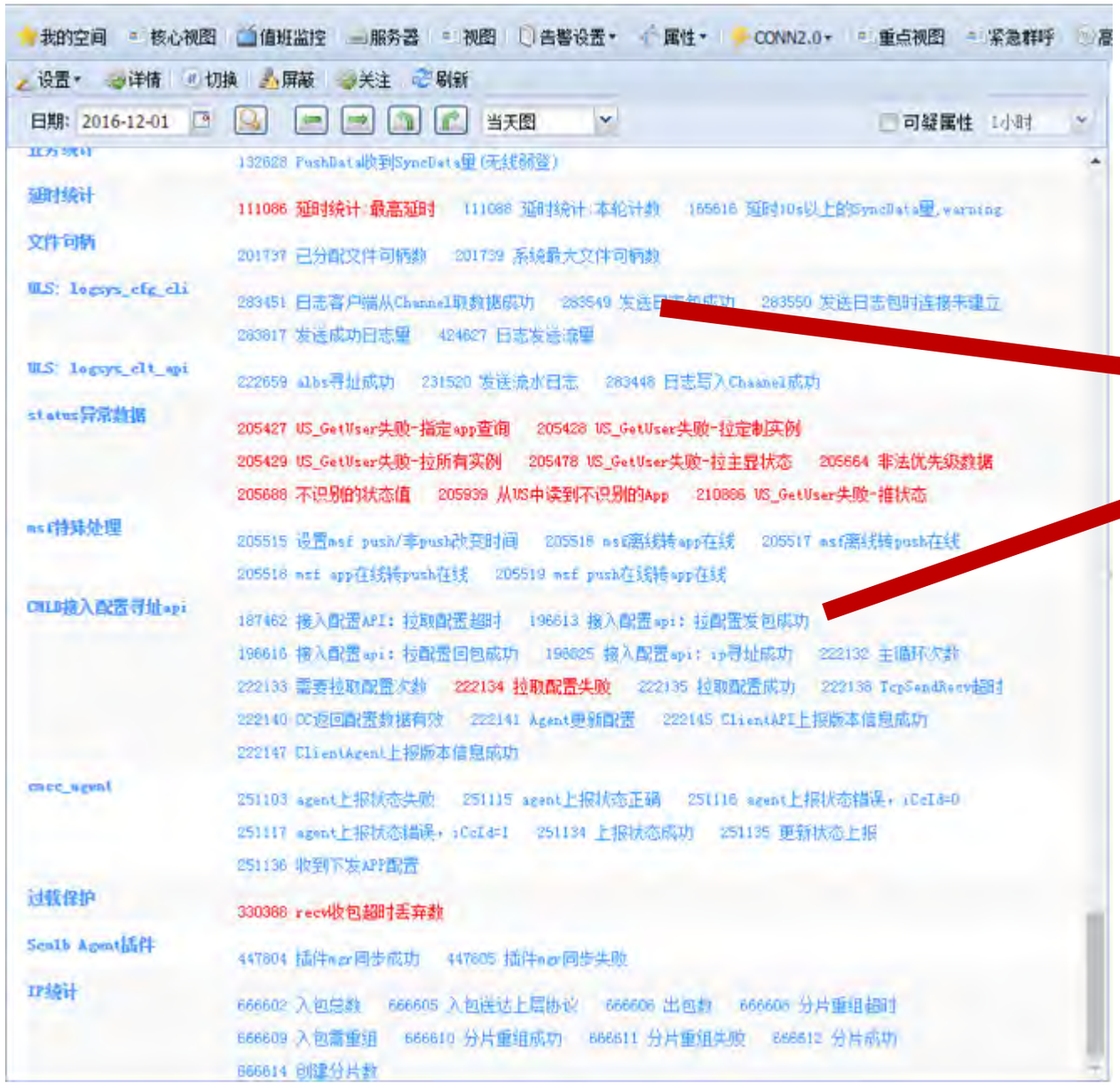
- 无阀值设定，完全根据指标值做波动判断
- 一个服务只能有一个生死指标
- 不建议用业务指标做生死指标

自定义-告警泛滥的罪魁祸首



无阀值设定，完全根据指标值做波动判断

一个服务只能有一个生死指标



DLP

案例跟踪

关联计算

自身故障/变更
网络故障公告
DBC故障公告

关联链路告警

全链路监控

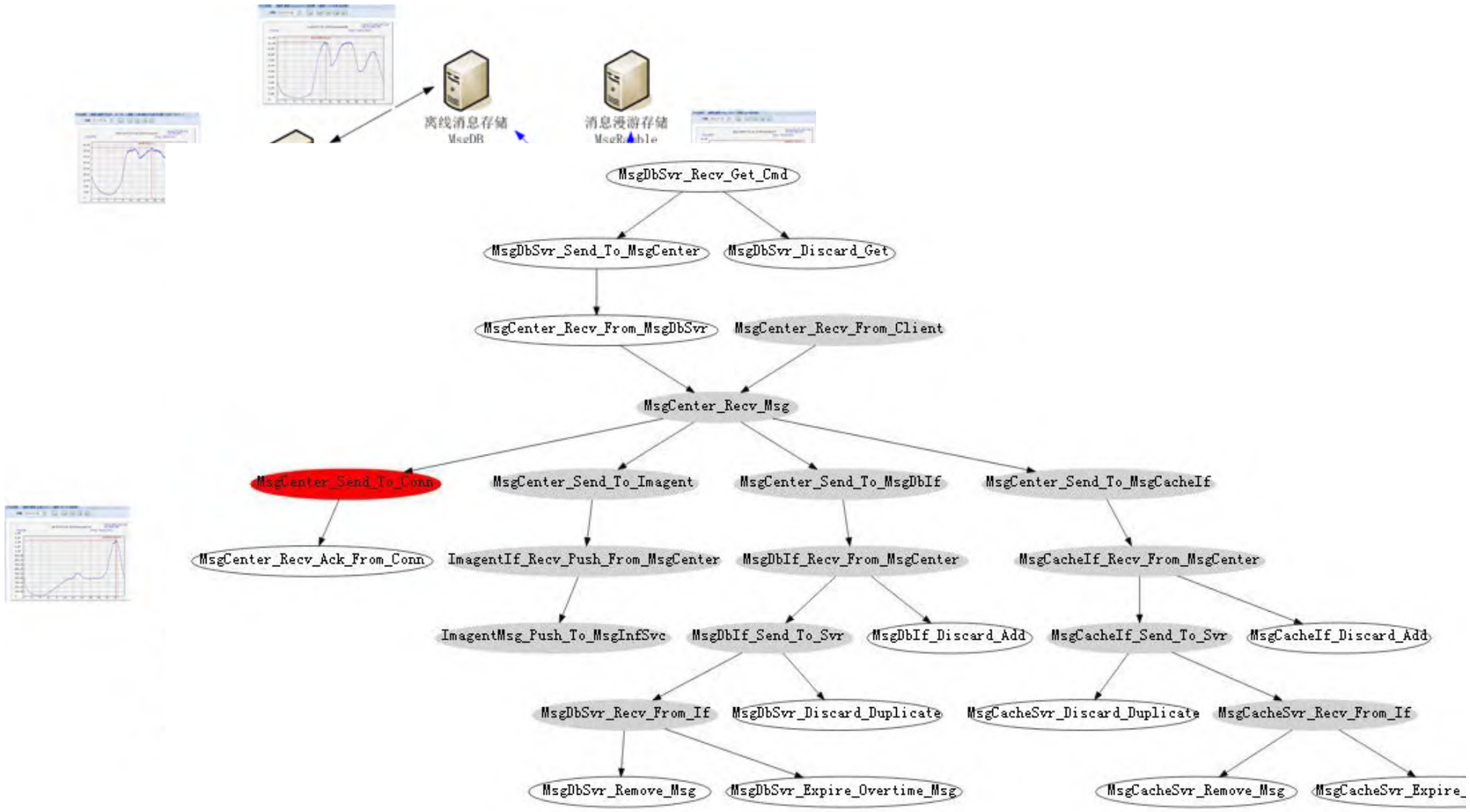
帮业务组织数据

腾讯织云
CLOUD
MANAGEMENT

无接入门槛的数据组织方式

- 不把数据多当包袱
- 界定数据的生产者 与 消费者
- 帮助生产者消费数据

早期染色监控



一条消息在系统内的51个状态

兼容各种数据源

织云基础监控
织云特性监控
其他已有数据源

织云多维监控

现网各类日志

织云日志中心

业务格式数据
织云舆情监控

筛选 & 转意预处理

抽取公共特性

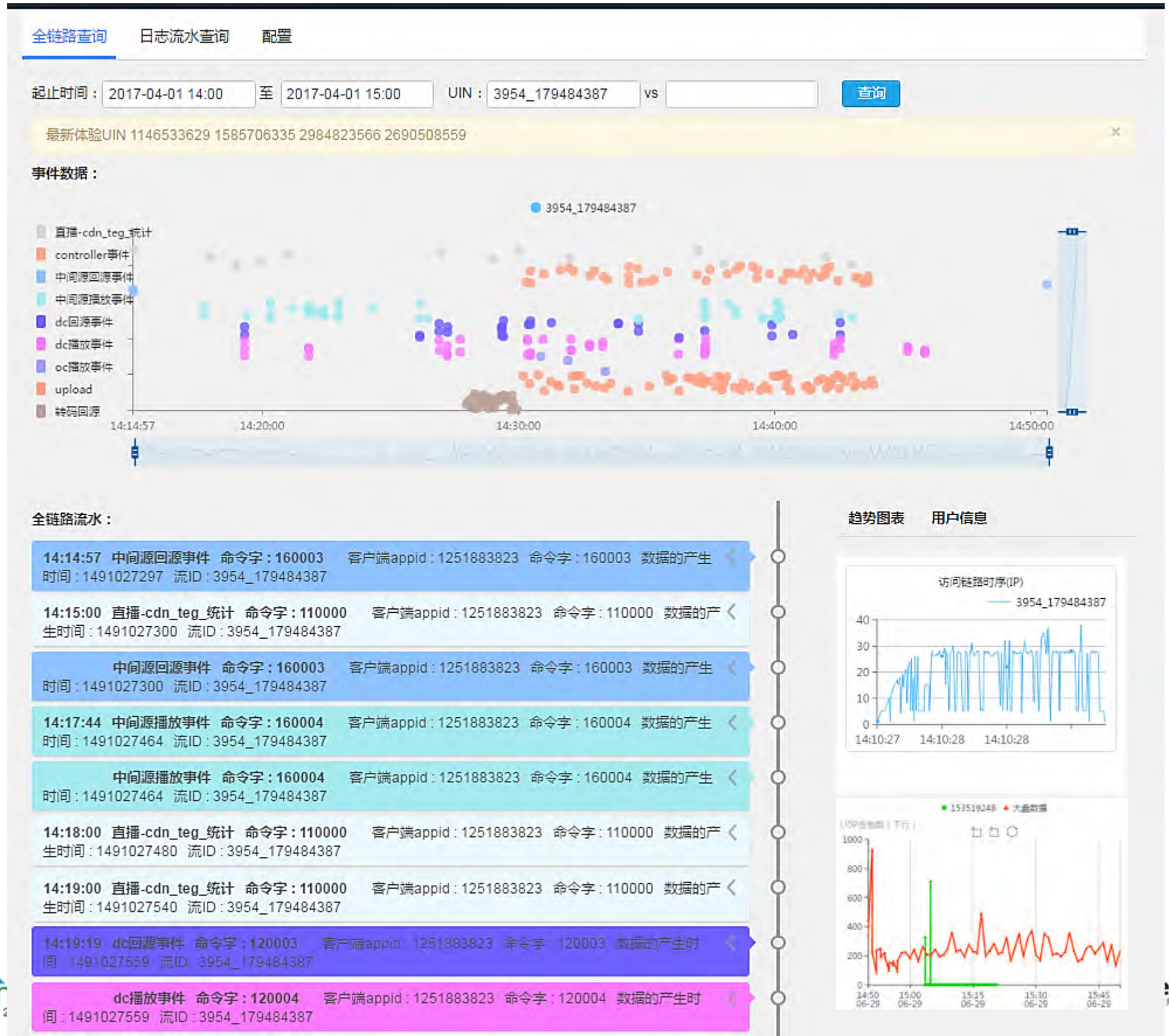
预计算大盘值

Elasticsearch

展现 & 告警

全链路分析监控平台

各类数据的再利用


全链路查询
日志流水查询
配置
起止时间：2017-04-01 14:00
至
2017-04-01 15:00
UIN：3954_179484387
vs
查询
最新体验UIN 1146533629 1585706335 2984823566 2690508559
事件数据：
3954_179484387
直播-cdn_teg_统计
controller事件
中间源回源事件
中间源播放事件
dc回源事件
dc播放事件
oc播放事件
upload
转码回源
14:14:57
14:20:00
14:30:00
14:40:00
14:50:00
全链路流水：
14:14:57 中间源回源事件 命令字：160003 客户端appid：1251883823 命令字：160003 数据的产生时间：1491027297 流ID：3954_179484387
14:15:00 直播-cdn_teg_统计 命令字：110000 客户端appid：1251883823 命令字：110000 数据的产生时间：1491027300 流ID：3954_179484387
中间源回源事件 命令字：160003 客户端appid：1251883823 命令字：160003 数据的产生时间：1491027300 流ID：3954_179484387
14:17:44 中间源播放事件 命令字：160004 客户端appid：1251883823 命令字：160004 数据的产生时间：1491027464 流ID：3954_179484387
中间源播放事件 命令字：160004 客户端appid：1251883823 命令字：160004 数据的产生时间：1491027464 流ID：3954_179484387
14:18:00 直播-cdn_teg_统计 命令字：110000 客户端appid：1251883823 命令字：110000 数据的产生时间：1491027480 流ID：3954_179484387
14:19:00 直播-cdn_teg_统计 命令字：110000 客户端appid：1251883823 命令字：110000 数据的产生时间：1491027540 流ID：3954_179484387
14:19:19 dc回源事件 命令字：120003 客户端appid：1251883823 命令字：120003 数据的产生时间：1491027559 流ID：3954_179484387
dc播放事件 命令字：120004 客户端appid：1251883823 命令字：120004 数据的产生时间：1491027559 流ID：3954_179484387
趋势图表
用户信息
访问链路时序(IP)
3954_179484387
40
30
20
10
0
14:10:27
14:10:28
14:10:28
153519248
大盘数据
1000
800
600
400
200
0
14:50
06-29
15:00
06-29
15:15
06-29
15:30
06-29
15:45
06-29

数据中挖掘各种纬度

数据源： 空间点播APP(dc00895)　查 询

成功率TOP10

事件	成功率	事件量	耗时
video_play_ret	99.96700	1409087485	0.000
video_downloa...	99.98400	1116061822	0.000
video_hotlink_...	99.98900	1382539093	0.388
video_downloa...	100.00000	220615050	0.000
video_link_redi...	100.00000	861936568	0.070
video_link_redi...	100.00000	867437443	0.000
video_downloa...	100.00000	1082542483	0.046

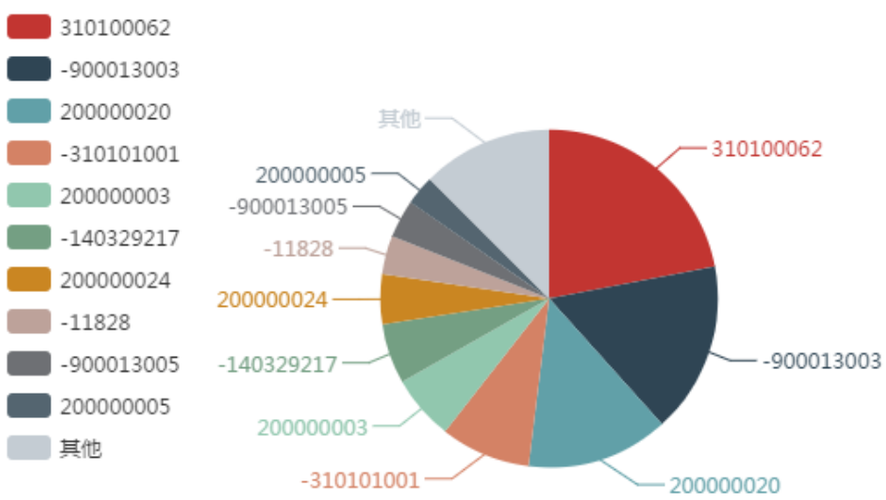
事件量TOP10

事件	成功率	事件量	耗时
video_play_ret	99.96700	1409087485	0.000
video_hotlink_...	99.98900	1382539093	0.388
video_downloa...	99.98400	1116061822	0.000
video_downloa...	100.00000	1082542483	0.046
video_link_redi...	100.00000	867437443	0.000
video_link_redi...	100.00000	861936568	0.070
video_downloa...	100.00000	220615050	0.000

耗时TOP10

事件	成功率	事件量	耗时
video_hotlink_...	99.98900	1382539093	0.388
video_link_redi...	100.00000	861936568	0.070
video_downloa...	100.00000	1082542483	0.046
video_downloa...	99.98400	1116061822	0.000
video_downloa...	100.00000	220615050	0.000
video_link_redi...	100.00000	867437443	0.000
video_play_ret	99.96700	1409087485	0.000

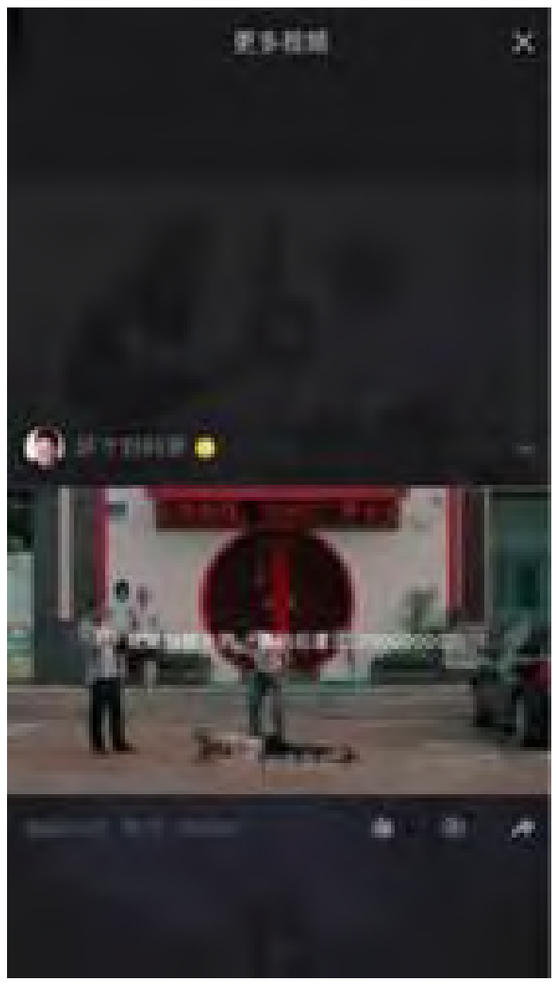
事件video_play_ret异常返回码占比分析



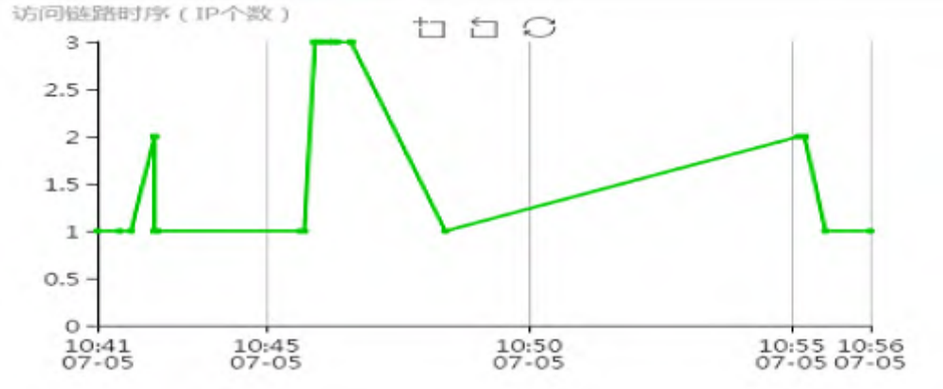
返回码	大盘占比	返回码描述
310100062	0.2195475	视频状态不合法
-900013003	0.1638776	播放器加载超时
200000020	0.1360407	socket错误
-310101001	0.0863267	http连接超时
200000003	0.0618855	DNS解析失败
-140329217	0.0578005	vkey校验失败
200000024	0.0477247	发情请求错误

举个栗子吧

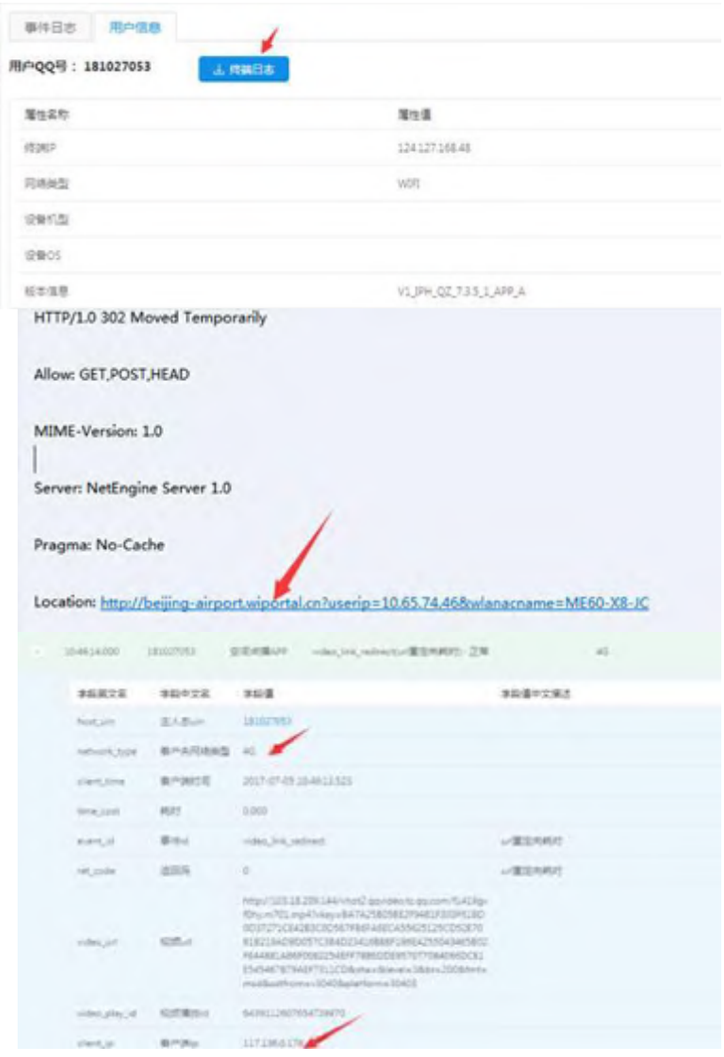
收到用户反馈



织云查该日志



验证分析结果



至此：根因是用户进了机场后，因wifi开关开启，自动试连，并跳转登陆界面

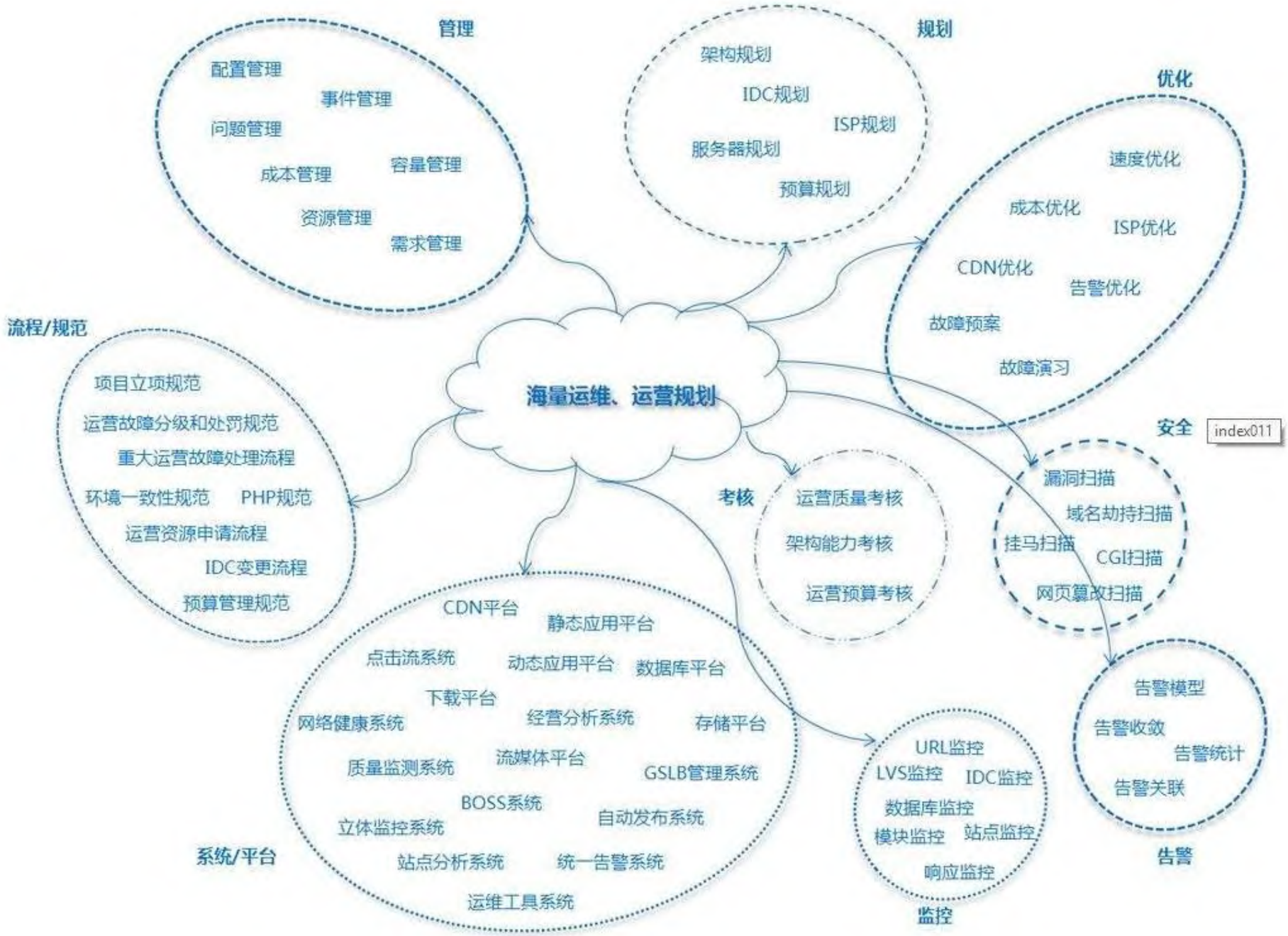
跟进时代，践行机器学习

CLOUD
MANAGEMENT

- 海量业务的监控**优势**
- 机器该学习**什么**
- **教**机器正确学习

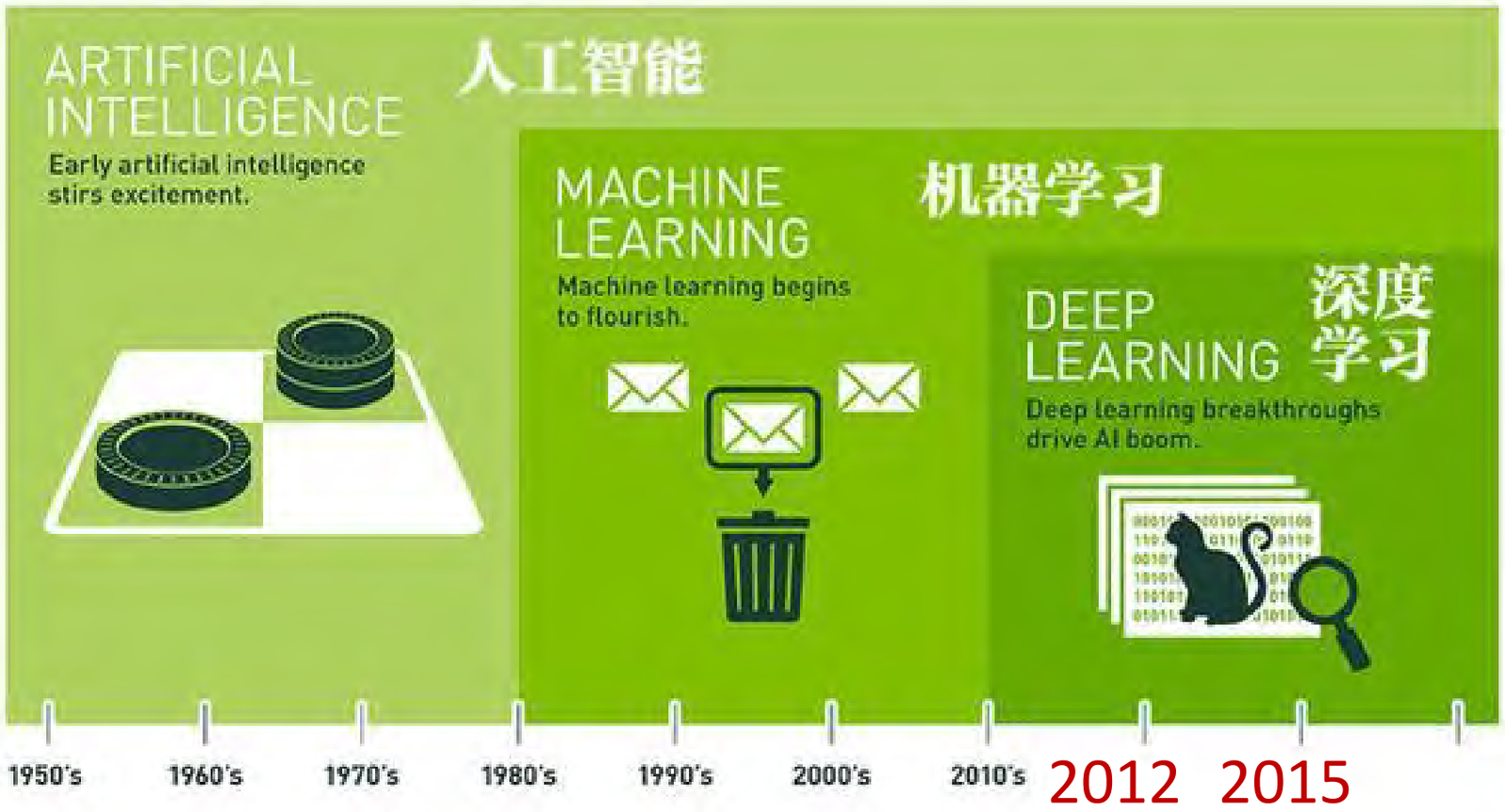
咖啡运维

运维需要这么多技能吗


海量运维、运营规划
管理
配置管理
事件管理
问题管理
成本管理
容量管理
资源管理
需求管理
规划
架构规划
IDC规划
ISP规划
服务器规划
预算规划
优化
速度优化
成本优化
ISP优化
CDN优化
告警优化
故障预案
故障演习
安全
index011
漏洞扫描
域名劫持扫描
挂马扫描
CGI扫描
网页篡改扫描
流程/规范
项目立项规范
运营故障分级和处罚规范
重大运营故障处理流程
环境一致性规范
PHP规范
运营资源申请流程
IDC变更流程
预算管理规范
考核
运营质量考核
架构能力考核
运营预算考核
系统/平台
CDN平台
静态应用平台
点击流系统
动态应用平台
数据库平台
下载平台
网络健康系统
经营分析系统
存储平台
质量监测系统
流媒体平台
GSLB管理系统
BOSS系统
立体监控系统
自动发布系统
站点分析系统
统一告警系统
运维工具系统
监控
URL监控
LVS监控
IDC监控
数据库监控
模块监控
站点监控
响应监控
告警
告警模型
告警收敛
告警统计
告警关联

AI 走向咖啡运维之路

狭义人工智能。对于特定任务，这样的技术能做得像人类一样好，甚至更好



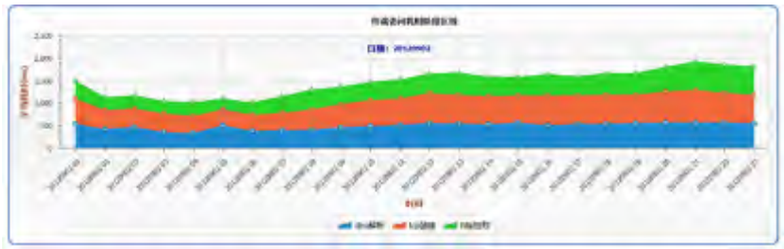
视频

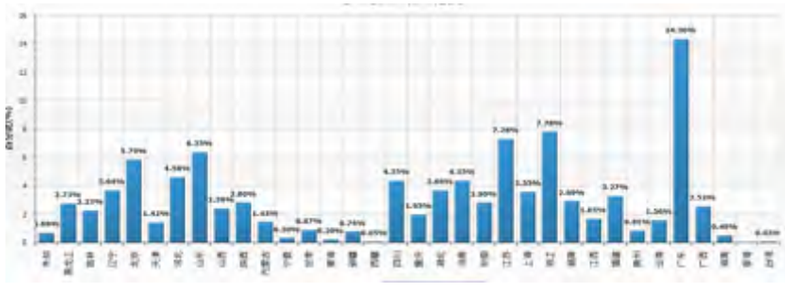
图像

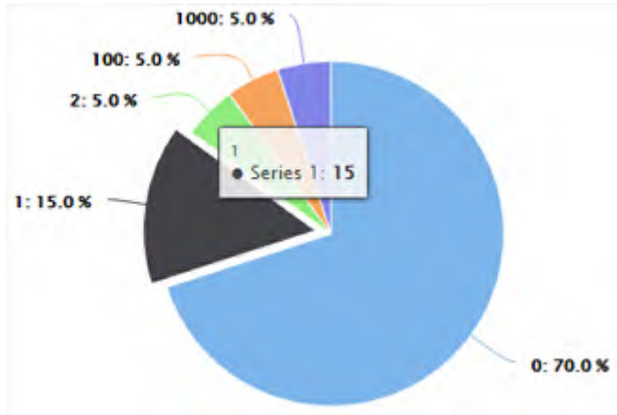
文本

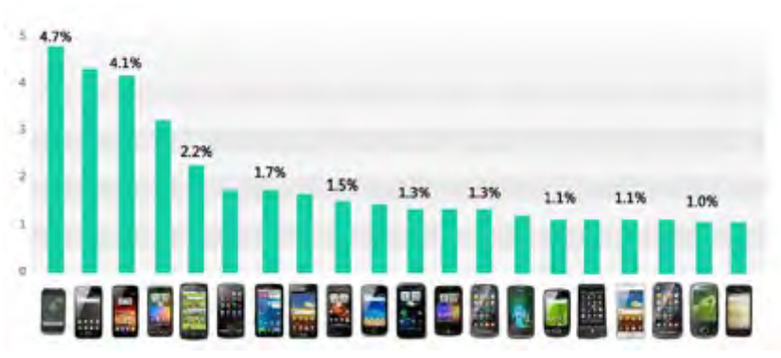
推荐

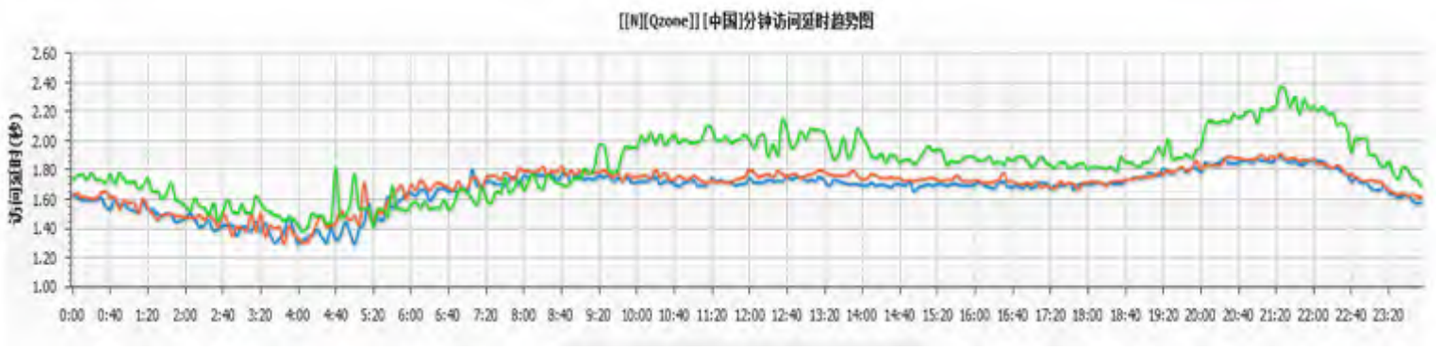
常见的分析模型











趋势、对比、波动、阀值、分布、聚类

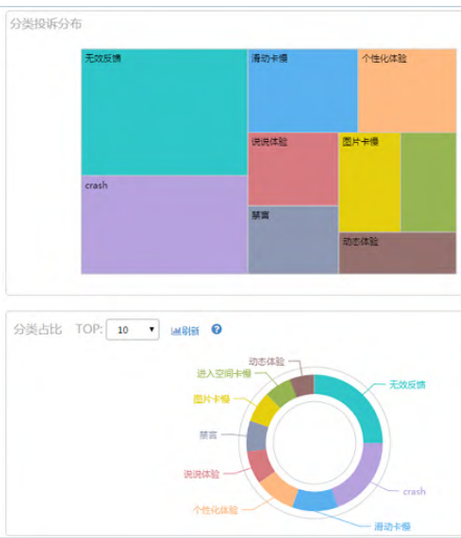
重新检视ROOT、DLP、全链路

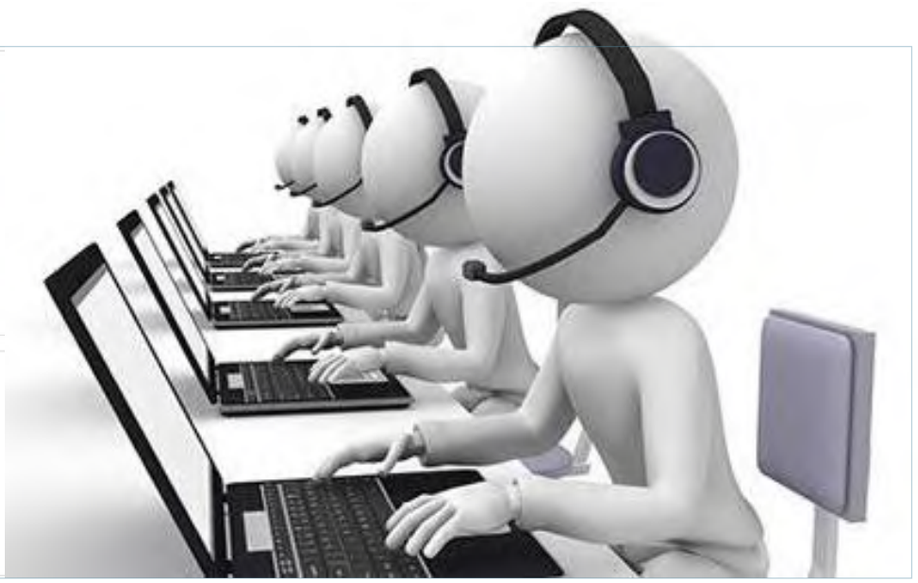
ROOT	基于架构 基于经验 基于概率	▶	*收敛告警事件*
DLP	基于规范 基于分工	▶	*产生告警事件*
全链路	基于数据 基于模型	▶	*提高事件处理能力*

第一个阶段　　机器学习 之 文本

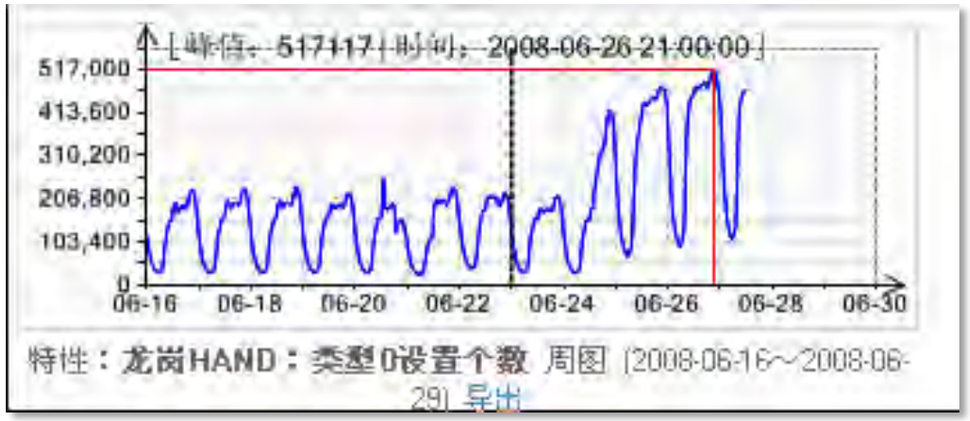
数据积累 → 问题发现 → 问题处理

织云舆情监控 + AI客服

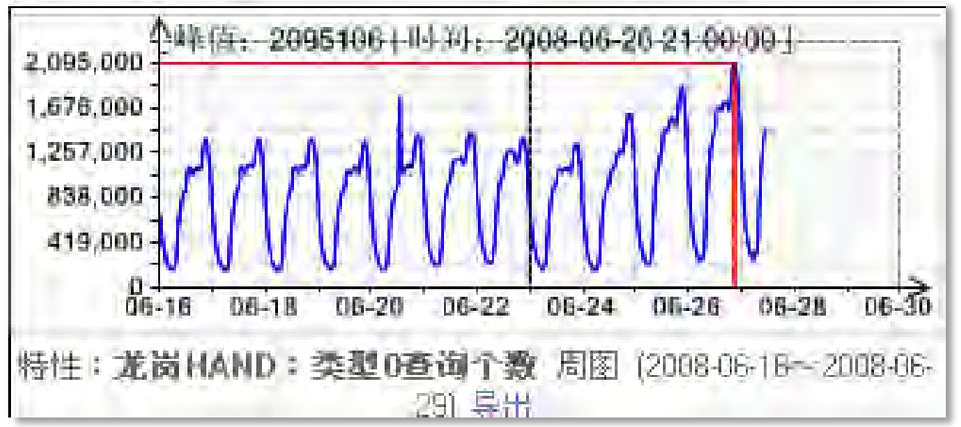




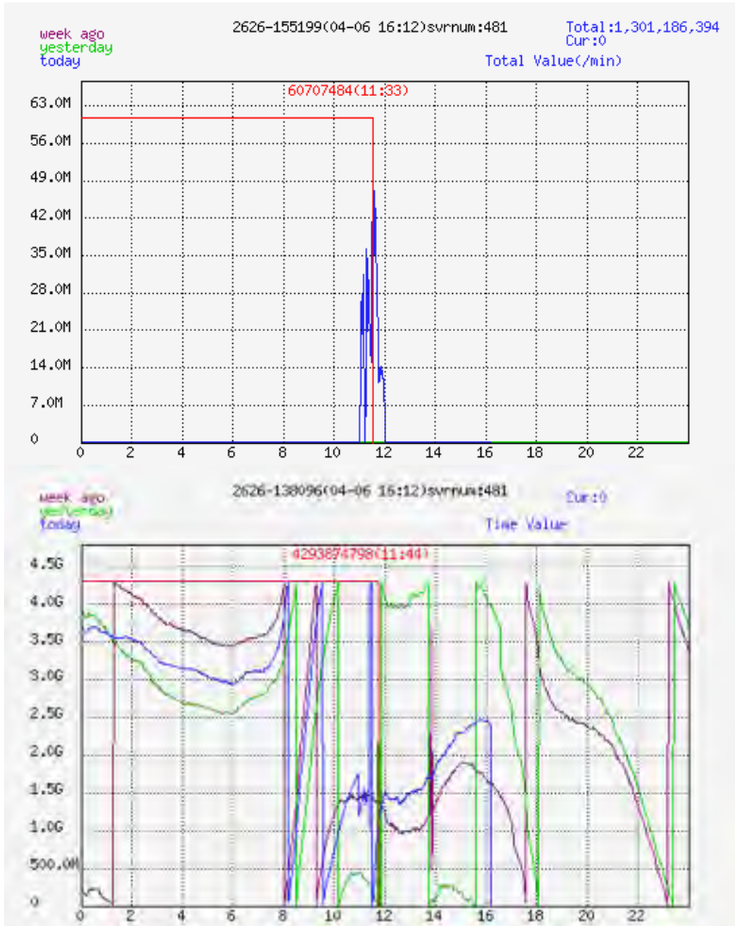
第二个阶段　　机器学习 之 图像



图像相似性



第三个阶段　如何告诉AI规则是什么

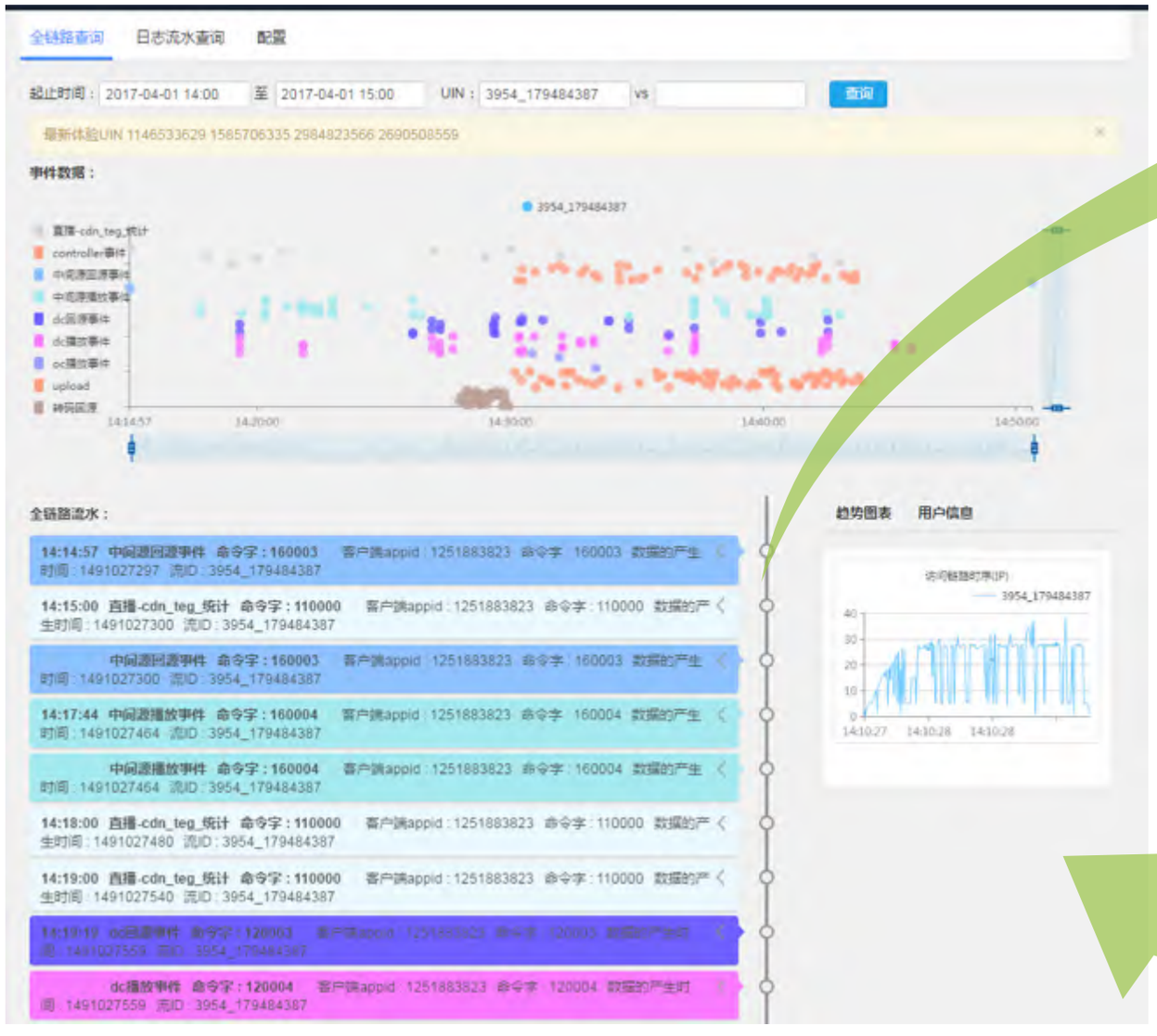


有监督学习

告警

如果用了自动找出来的历史告警，模型就会学成历史告警的策略 = nothing

第四个阶段　告诉AI数据的意义



机器学习

有哪些有值得关注点

腾讯织云
CLOUD
MANAGEMENT

监控是 平台
也是 产品
更重要是 运营

如果监控是产品

快 即时性 告警快

准 告警准 误告少

全 无遗漏 覆盖广

如果监控是平台

稳

强

易

如何运营监控

指标

闭环

生态

指标

SMART

- 很具体
- 可衡量
- 可达到
- 可观察
- 有时间

DLP告警 = 服务异常

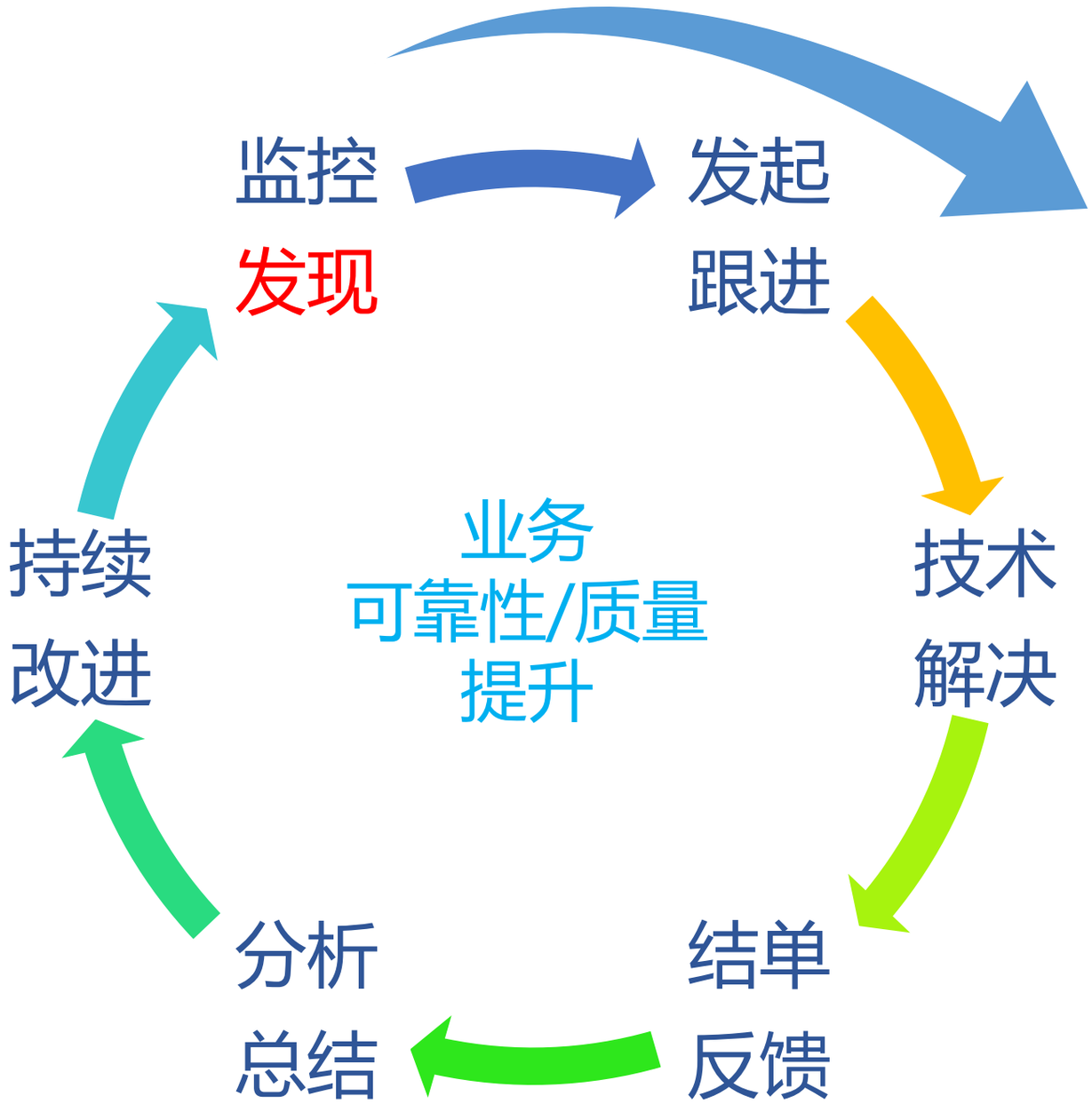
服务异常 x DLP告警时间 = 服务不可靠性

SUM（服务可靠性,……） = 业务不可靠性

业务不可靠性/一段时间(如月,年) = 业务可靠性**指标**(年,月)

- 可横向对比
- 有趣势变化
- 可以目标明确

闭环



生态

业务 

用户端卡慢

自动化测试

摸调

摸调

业务 

舆情监控

自动化测试

摸调

摸调

业务 

H5/测速

自动化测试

摸调

摸调

Crash /卡慢
- 移动端监控

速度
- APP H5测速
- web测速

体验
- 多媒体图片
- 海外速度
- 运营商

成功率
- ATT
- 摸调
- monitor
- 业务特性告警
- 组件监控

体验有提升

成功率改进

可用性衡量

横向业务**指标**对比

分析手段

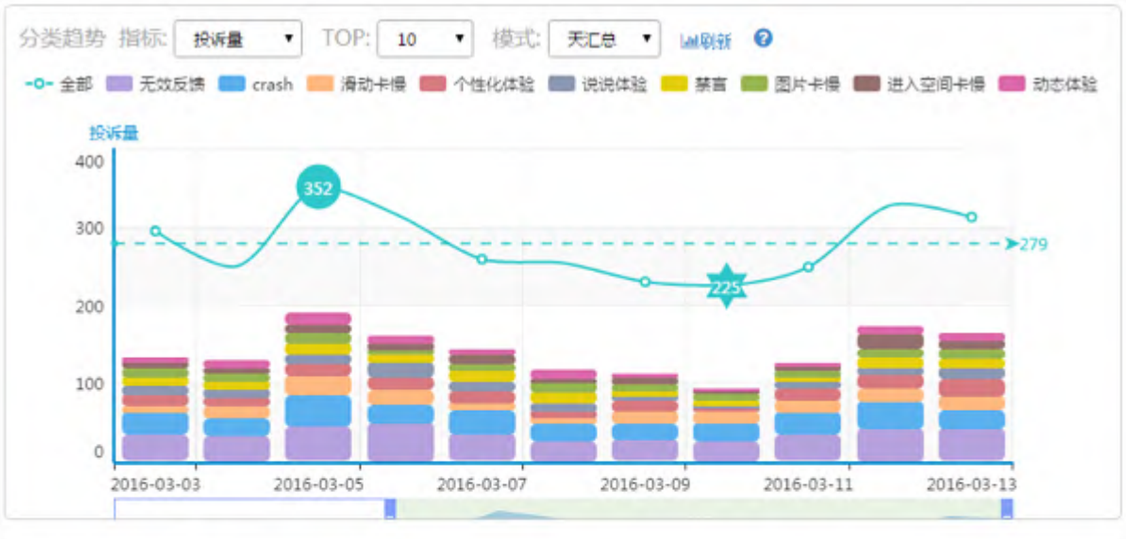
目标

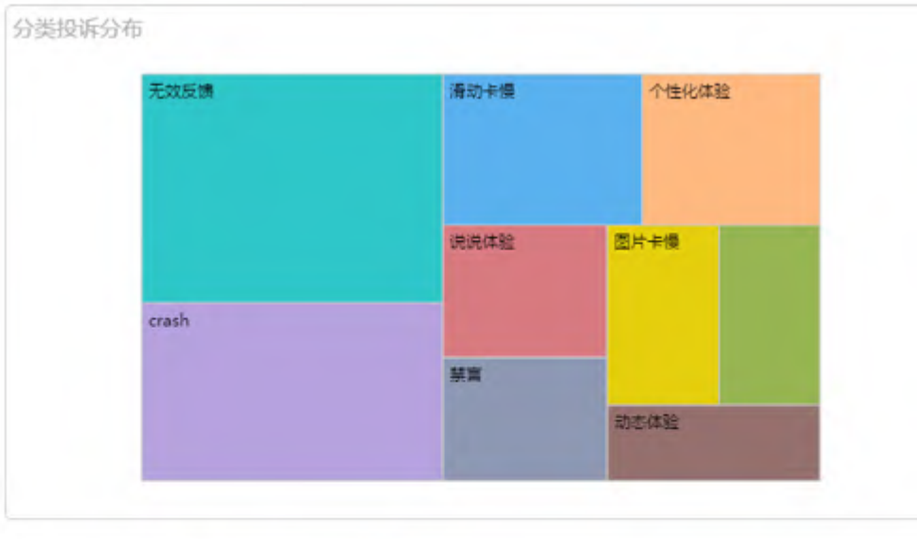
TIPS

腾讯织云
CLOUD
MANAGEMENT

自动处理、
移动办公、
舆情监控、
告警分级

织云舆情监控

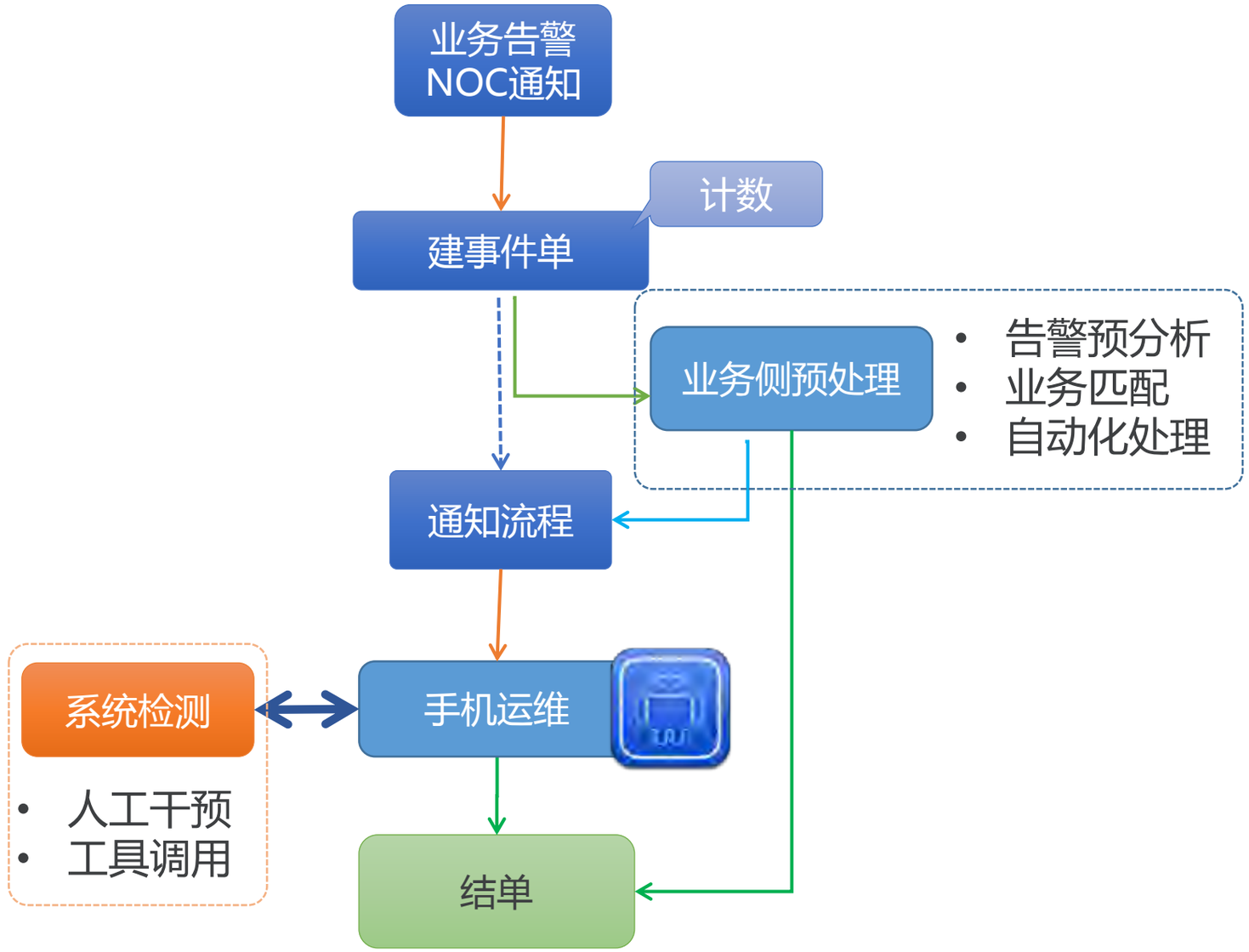

分类趋势 指标: 投诉量 TOP: 10 模式: 天汇总 刷新
全部 无效反馈 crash 滑动卡慢 个性化体验 说说体验 禁言 图片卡慢 进入空间卡慢 动态体验
投诉量
400
300
200
100
0
352
279
225
2016-03-03
2016-03-05
2016-03-07
2016-03-09
2016-03-11
2016-03-13


分类投诉分布
无效反馈
滑动卡慢
个性化体验
crash
说说体验
图片卡慢
禁言
动态体验


分类-版本投诉 全部 TOP: 5 刷新
400
300
200
100
0
2016-02-28
2016-03-04
2016-03-09
6.2.3.409
未分类/不详-版本
5.9.5.451
6.2.2.402
6.1.0.496


分类占比 TOP: 10 刷新
动态体验
进入空间卡慢
图片卡慢
禁言
说说体验
个性化体验
无效反馈
crash
滑动卡慢

告警自动告警处理


业务告警
NOC通知
计数
建事件单
业务侧预处理
• 告警预分析
• 业务匹配
• 自动化处理
通知流程
系统检测
手机运维
• 人工干预
• 工具调用
结单

织云移动运维办公


中国移动
20:39
35%
主页
手机运维
返回
错误率
100.00 %
2 次
1 次


中国移动
20:39
35%
主页
手机运维
返回
自动化监控V4.0
屏蔽
GET /cgi-bin/yanchu/soso/soso_search.fcg
#GET /cgi-bin/yanchu/soso/soso_search.fcg
#:错误率100%&&错误数10
johnnyliu
kenvinma
babychen
jackytan
yahayapeng
yahayapeng
[消息通知服务][SET_SMSTIPS]
[逻辑SPP]
10
0.00 %
100.00 %
100.00 %


中国移动
20:39
35%
主页
手机运维
返回
自动化监控V4.0
屏蔽
GET /cgi-bin/yanchu/soso/soso_search.fcg
#GET /cgi-bin/yanchu/soso/soso_search.fcg
#:错误率100%&&错误数10
完成
屏蔽
屏蔽1天
屏蔽3天
结束屏蔽

监控如何分级

告警级别

时间分级

范围分级

推送规范

请交流，请携手