



主办方: msup | ARCHNOTES 架构
主办方: msup | ARCHNOTES

GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE

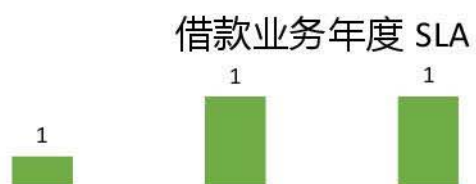
AIOps 根因溯源产品 在互联网的落地实践

朱颖航 Linkedsee 灵犀

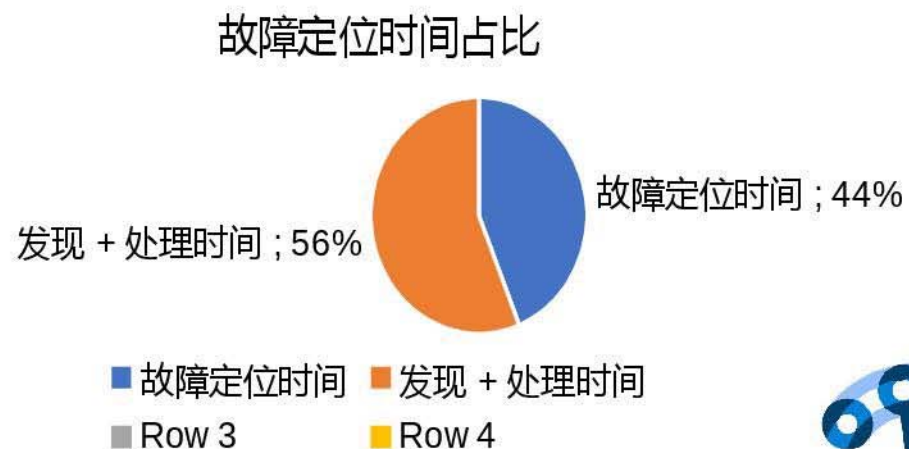
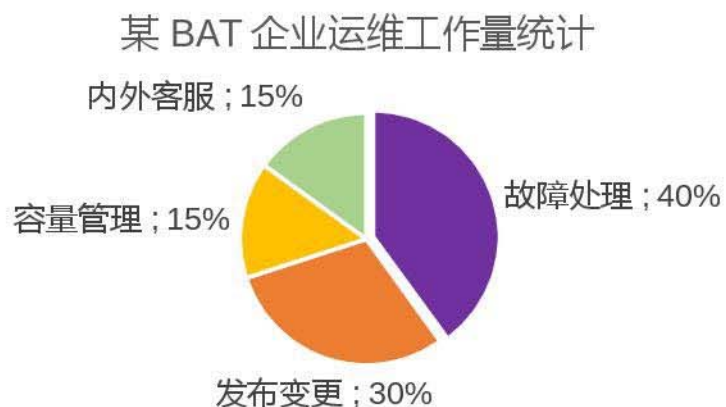
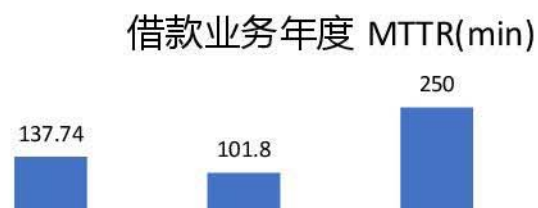


SLA 仍然是运维 No.1 KPI

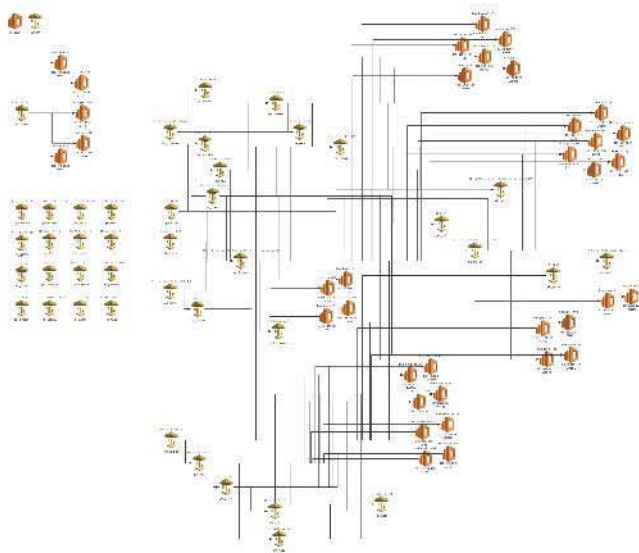
- 投入最大; 且始终充满挑战



- 故障排查仍然是 MTTR 最大瓶颈

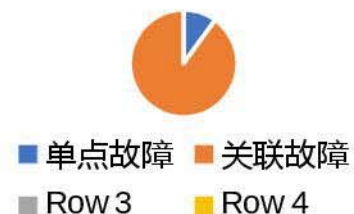


故障现状越来越复杂



- IT 基础设施是一套复杂的系统
- 这个系统呈齿轮状运转，相互依存
- 已经几乎不存在单点型故障

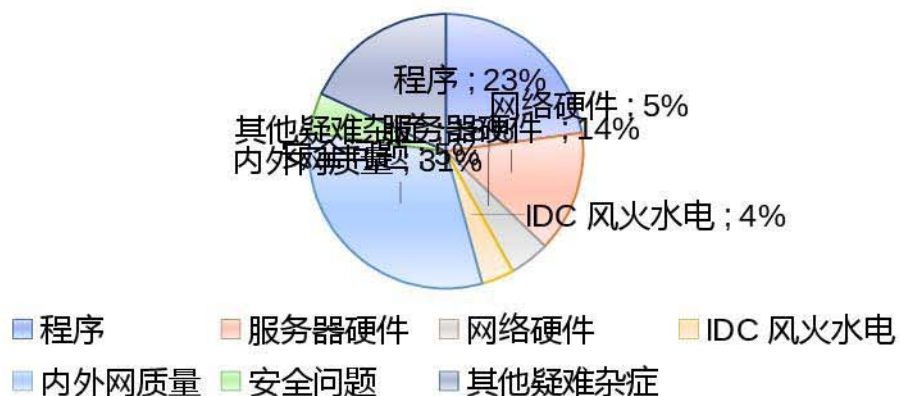
故障关联性占比



故障原因占比 - 概要

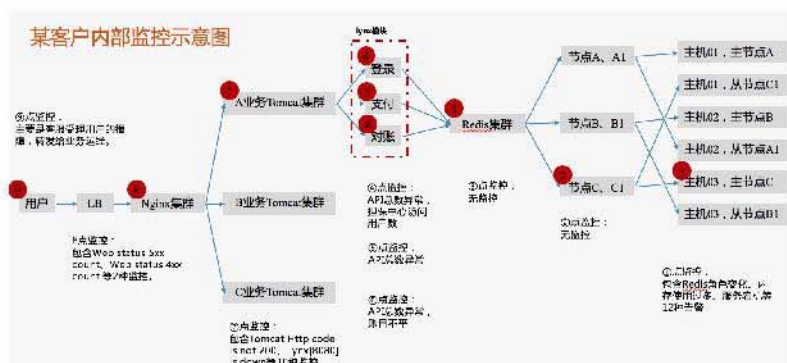


故障原因占比 - 具体

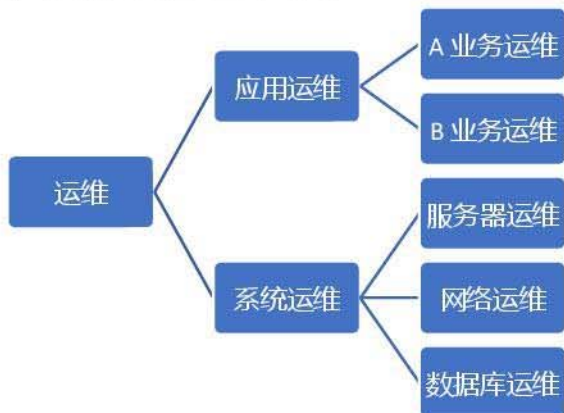


通常运维组织解决问题的思路

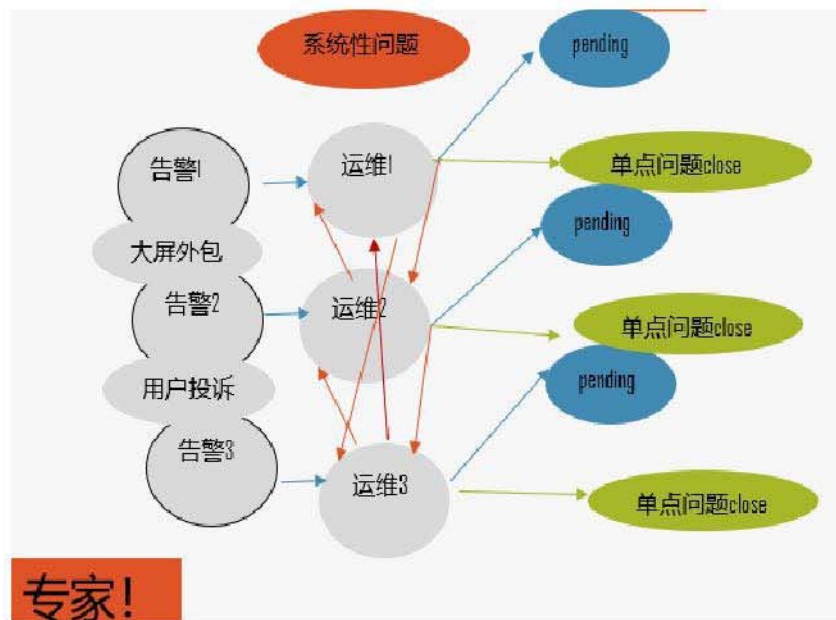
1、长期持续监控建设



2、不断细化的组织分工



3、交互频繁的团队协作



现状痛点 1 : 海量的告警 + 告警噪音 = 大量被浪费的人力

告警列表

告警ID	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0001	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0002	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0003	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0004	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0005	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0006	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0007	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0008	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0009	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0010	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0011	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0012	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0013	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0014	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0015	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0016	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0017	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0018	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0019	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0020	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0021	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0022	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0023	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0024	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0025	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0026	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0027	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0028	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0029	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人
告警-0030	告警名称	告警类型	告警时间	告警级别	告警内容	告警状态	告警处理人

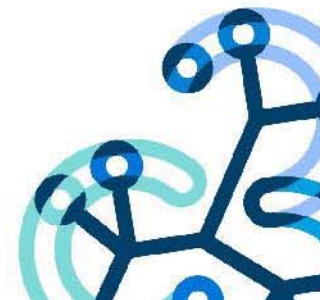
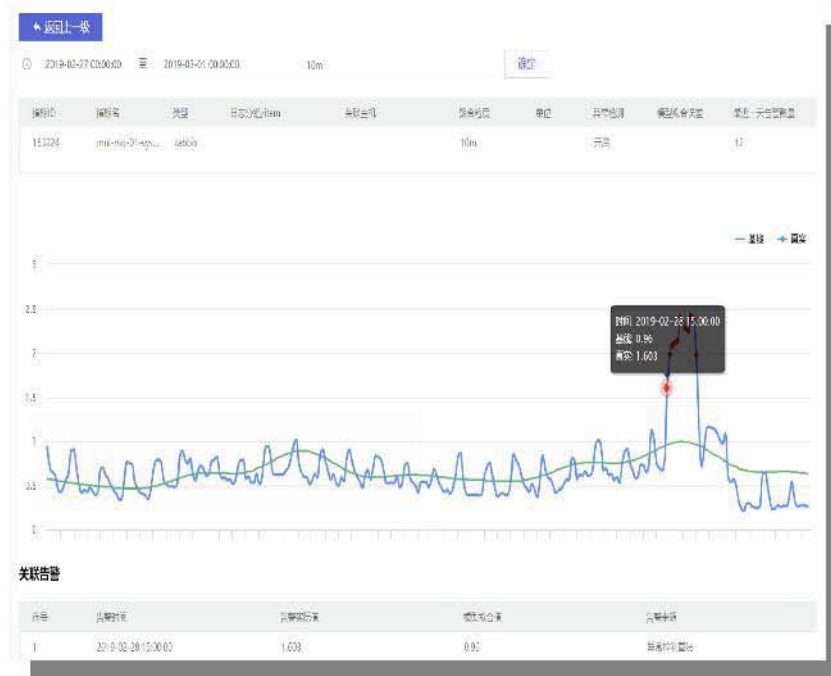
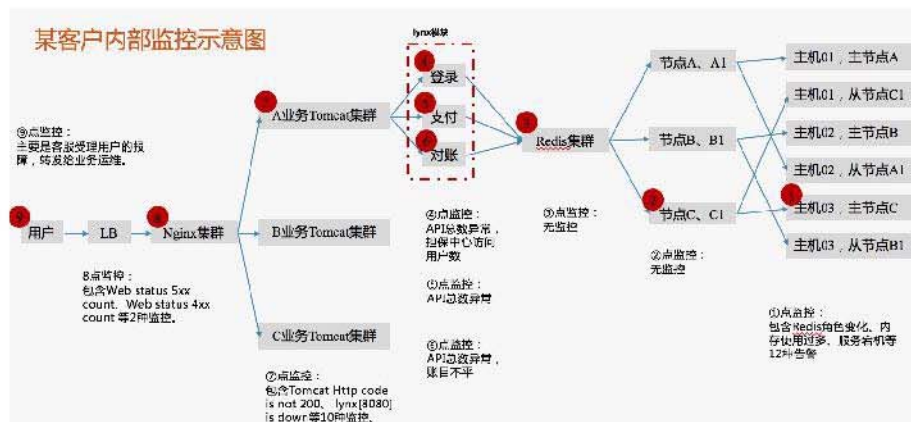
找到交易失败对应的日志, 根据交易id找到交易请求日志, 从中自动解析客户信息

关联该用户最近三十分钟做过的其他交易, 如果包含大量失败交易, 可能是该用户状态出现异常或者在发起攻击; 该故障中客户的几次请求中只失败了这一次, 无需技术人员过多关注

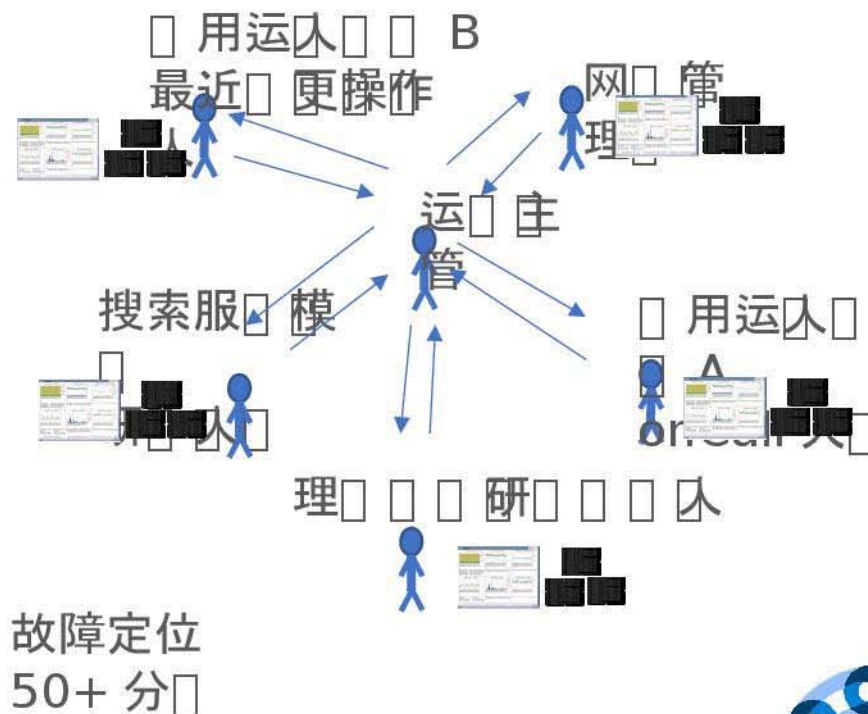
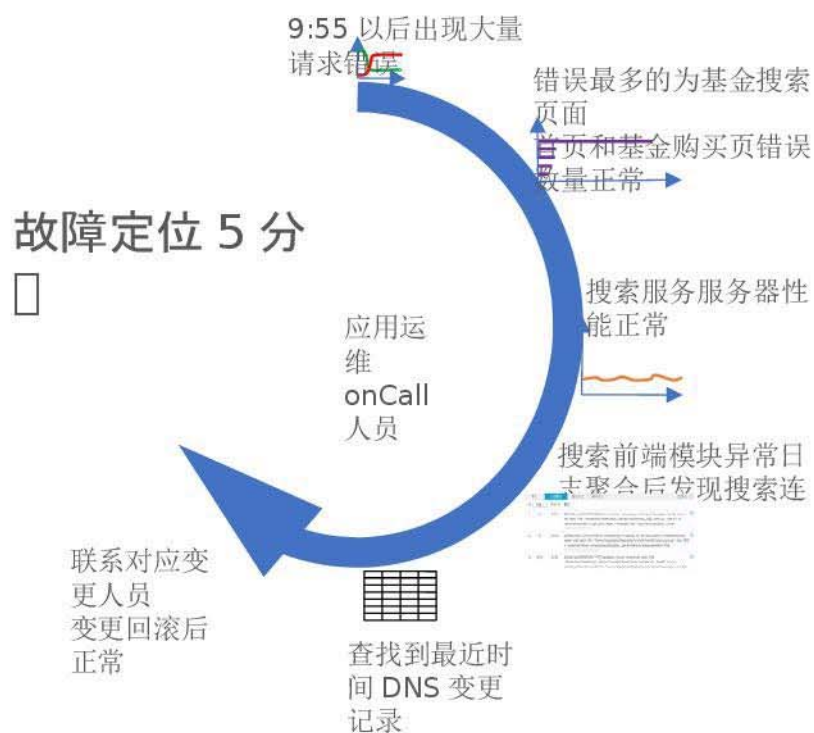
序号	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
1	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
2	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
3	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
4	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
5	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
6	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注
7	交易ID	交易时间	交易状态	交易内容	交易结果	交易备注



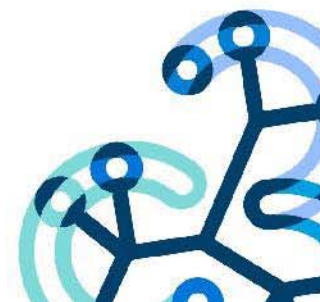
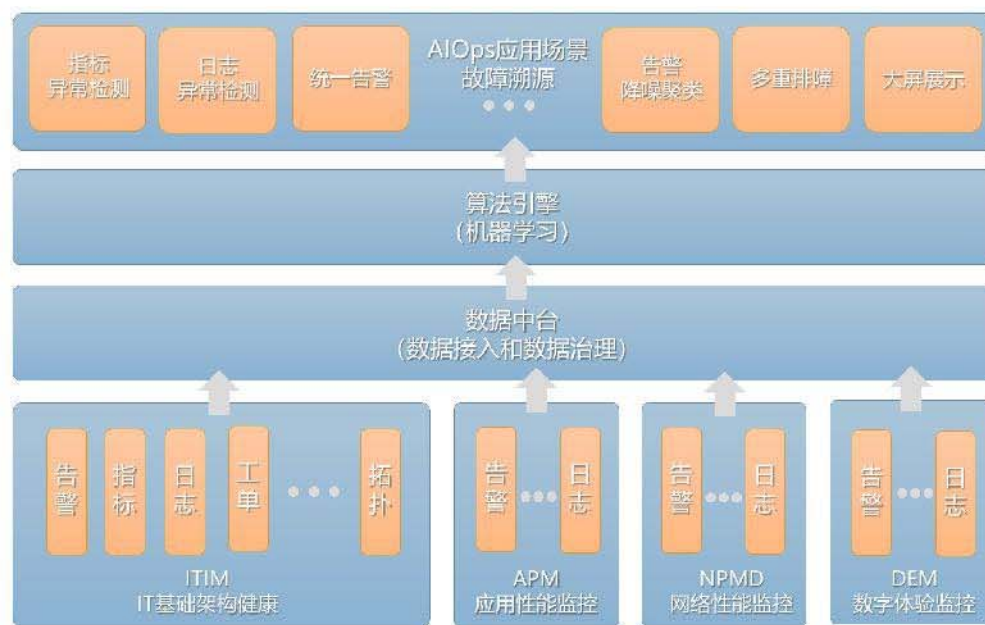
现状痛点 2 : 消失的告警 = 大幅滞后的故障发现时间



现状痛点 3 : 多人协同 + 专家依赖 = 排查问题又幸运又低效



LinkedAIOps 数据源



1 **指标异常检测**
通过算法自动学习和建立动态基线, 提前感知和发现异常

2 **日志异常检测**
通过算法自动分析业务运行日志异常, 提前感知和发现潜在故障

3 **告警降噪聚类**
基于算法的告警降噪和聚类, 实现 90% 以上的降噪效果



多重故障溯源
通过算法将包括指标异常检测、日志异常检测、故障根因推荐等多种数据关联, 实现快速故障定位

一键排障
通过最简单的一键排障入口, 实现最高效的故障排查

大屏展示
通过直观的关键运维分析数据的展示, 实时掌控关键业务的运行健康状态



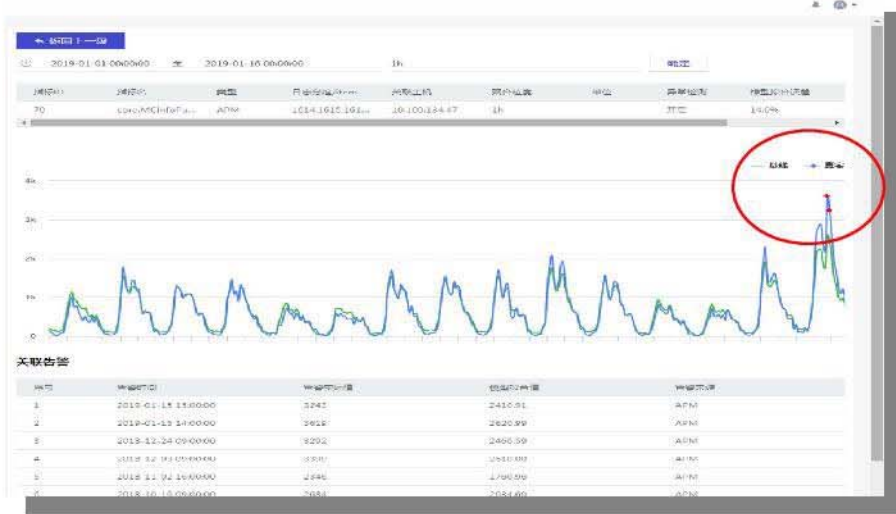
产品功能 - 统一数据接入和治理

数据源	支持的对接方式	客户若当前无此数据源	实时性	是否必需接入
系统告警 (含 IAAS+PAAS 层)	开源工具 (Zabbix 、 Open - falcon 、 Prometheus) 、 Pinpoint APM 、 API	可推荐行业内厂商；或帮助客户部署	秒级	是
业务告警	BPM 、 API	可通过日志数据帮客户梳理	秒级	是
日志	ELK 、 Rsyslog 、 Syslog	可推荐行业内厂商；或帮助客户部署	分钟级	否
指标 (含业务指标和系统指标)	开源工具 (Zabbix 、 Open - falcon) 、 API	可帮助客户梳理	分钟级	否
工单	Jenkins 、 ITSM 、 API	暂无推荐和代为部署能力	分钟级	否



产品功能 - 指标异常检测

- 使用基于深度学习的异常检测算法，替代传统基于固定阈值的监控方案
- 特点:
 - 学习历史数据，分析当前指标曲线趋势是否异常
- 优势:
 - 零阈值，不再需要配置阈值
 - 告警准确率高
 - 更早发现异常情况
 - 可适应业务发展带来的趋势变化



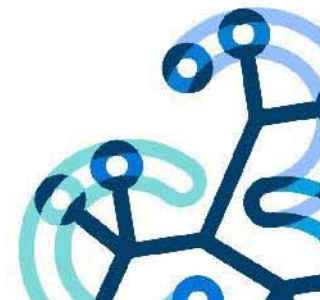
产品功能 – 日志异常检测

- 对故障时段的日志做聚类形成日志模式，并与正常时段的日志模式做对比，发现异常日志行为，定位故障原因
- 价值：
 - 出现故障后迅速定位原因，减少 MTTR
 - 弥补现有监控规则覆盖不到的故障情况
- 算法特点：
 - 无需对日志做解析
 - 针对监控规则没覆盖的场景，价值大，准确率高

我的故障模式 所有故障模式(30) 未关闭的故障模式

紧急/正常/次要 主机 报警 全部状态 开始时间 结束时间 自定义 帮助

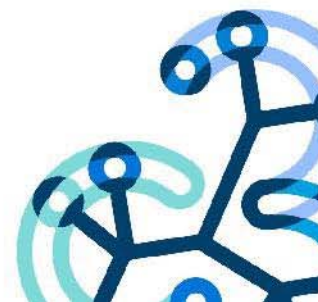
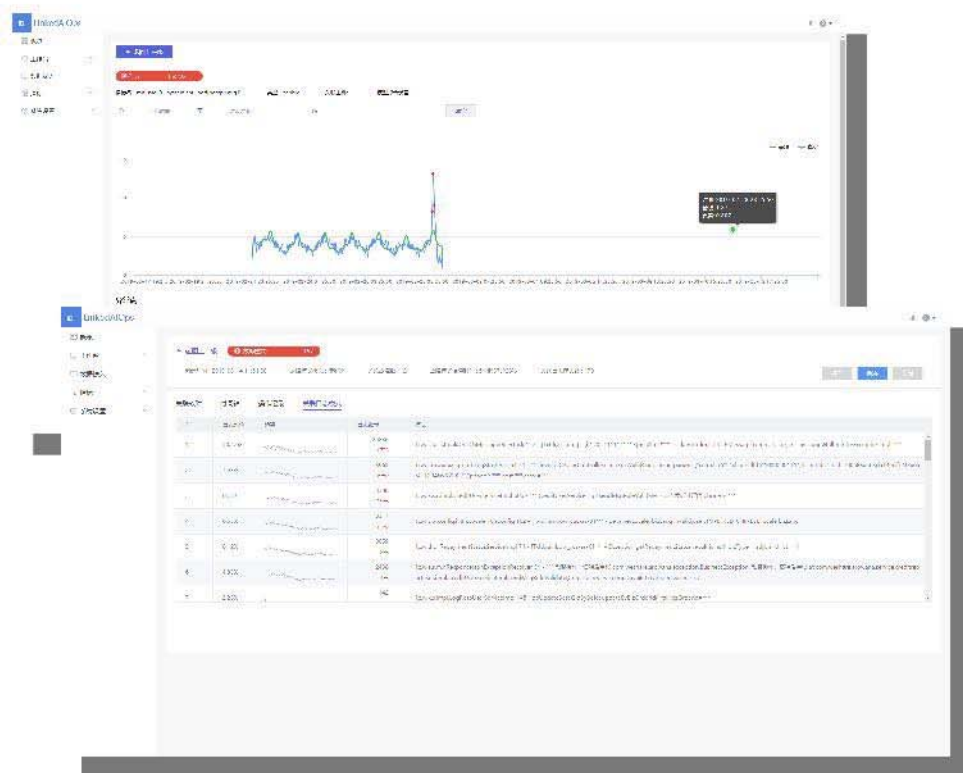
故障模式等级	ID	模式根因	创建时间	影响的主机数	影响的主机	关联故障数(50)
严重	1	数据异常日志模式: 数量:06	2019-04-12 22:20:00	1	—	2
严重	2	HTTP service port 143.16...	2019-04-12 22:27:01	1	UCM_10.100.15.01	2
严重	3	数据异常日志模式: 数量:19	2019-04-12 21:30:00	1	—	3
严重	4	主机 {HOST.NAME} 上的...	2019-04-12 21:07:10	1	SYQ080_10.100.1...	1
严重	5	主机 {HOST.NAME} 上的...	2019-04-12 19:48:01	1	wmln5750_10.10...	1
严重	6	数据异常日志模式: 数量:149...	2019-04-12 19:20:00	1	—	7
严重	7	数据异常日志模式: 数量:90...	2019-04-12 19:20:00	1	—	1
严重	8	主机 {HOST.NAME} 上的M...	2019-04-12 15:50:04	1	wmln7753_10.100...	1
严重	9	{HOST.NAME} Ping 失败...	2019-04-12 13:55:16	2	DLG-CNC-GW-1...	2



产品功能 - 告警降噪和辅助增强

- 基于统一数据接入后的统一告警中心
- 增加基于日志异常检测和指标异常检测的新告警

- 去重降噪, 将告警量减少 90%
- 异常检测算法, 将告警准确率提升至 70%



产品功能 - 告警聚合

- 将 1 个事实故障衍生的所有相关告警都聚类成 1 个故障
- 以此作为故障处理的第一站

[illegible]

- 直接推荐根因
- 关联日志、指标、调用栈、工单数据等辅助排查

返回上一级		故障模式		1793			
查看时间: 2019-03-21 12:08:00		故障模式预览		系统故障数: 11			
				故障模式总耗时: 5小时7分30秒			
				故障报警数: 10			
关联故障	时间轴	操作记录	关联指标	关联故障模式	Endpoint地址		
0	故障	4983	0.693	Tomcat mountain(0100) appver has nodata for 5 min.	cus-es-C1	Tomcat	2019-03-21 12:06:02
0	故障	4982	0.618	Tomcat rainbow(8090) appver has nodata for 5 min.	cus-es-C1	Tomcat	2019-03-21 12:06:03
0	故障	5068	0.125	Tomcat mountain(0100) appver has nodata for 5 min.	cus-es-C1b	Tomcat	2019-03-21 12:06:02
0	故障	5035	0.375	Tomcat mountain(8090) appver has nodata for 5 min.	cus-es-C2	Tomcat	2019-03-21 12:06:05
0	故障	5064	0.125	Tomcat rainbow(8090) appver has nodata for 5 min.	cus-es-C1c	Tomcat	2019-03-21 12:11:15
0	故障	5677	0.322	Tomcat rainbow(8090) appver has nodata for 5 min.	cus-es-C1b	Tomcat	2019-03-21 12:06:02
0	故障	5068	0.119	Tomcat rainbow(8090) appver has nodata for 5 min.	cus-es-C1d	Tomcat	2019-03-21 12:06:02
0	故障	5598	0.219	Tomcat poplar(8090) appver has nodata for 5 min.	cus-es-C3a	Tomcat	2019-03-21 12:06:02
0	故障	5084	0.216	Tomcat hall(0100) appver has nodata for 5 min.	cus-es-C2	Tomcat	2019-03-21 12:06:05
0	故障	5037	0.316	Tomcat lake(8280) appver has nodata for 5 min.	cus-es-C2	Tomcat	2019-03-21 12:06:05
0	故障	5099	0.21	Tomcat poplar(8090) appver has nodata for 5 min.	cus-es-C3	Tomcat	2019-03-21 12:06:02



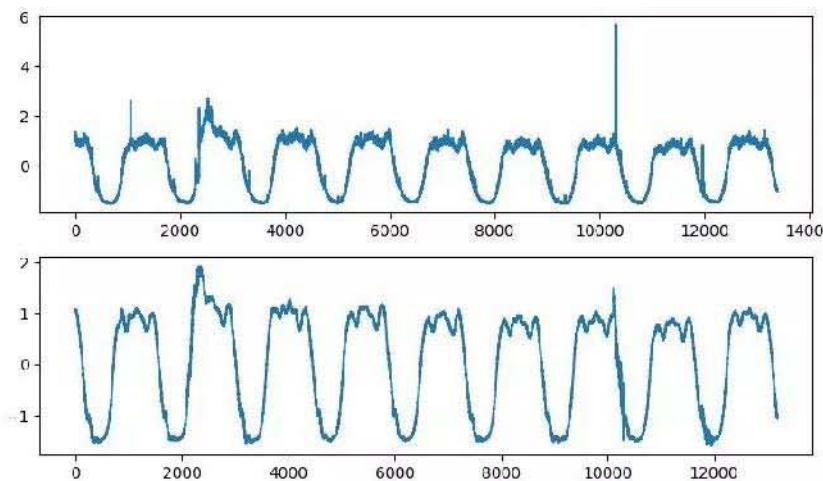
指标异常检测算法

- 发现系统和业务指标的异常情况（异常上涨 / 下跌）
- 难点：
 - 如何定义基线
 - 根据基线如何定义异常
- 基线定义方法：
 - 规则
 - 回归：基于时序分解



指标异常检测——基于时序分解确定基线

- 插值补缺
- 平滑降噪
- 局部加权非参数模型
- 季节性周期分解：年、季、月、周、日
- 差分滑动平均自回归模型



指标异常检测——无监督算法如何确定异常

- 基于统计学

- 基于分布的估计
- 对数似然估计
- 基于周期环比

效果：

召回率较高

准确率中等

- 基于机器学习：

- VAE
- 孤立森林
- EWMA
- 逻辑回归



指标异常检测——有监督异常检测

- 特征工程 (140+) :
 - 统计特征
 - 拟合特征
 - 分类特征
- 分类器 ($F1 > 0.65$) :
 - XGBoost
 - DNN
 - GBDT



指标异常检测——模型设计

- 在线模型:
 - 对接监控系统拉取最新数据 (延迟 $< 10\text{min}$)
 - 实时判断是否异常 (延迟 $< 10\text{s}$)
- 离线模型:
 - 对每条曲线重新训练和定阶 (每半小时)



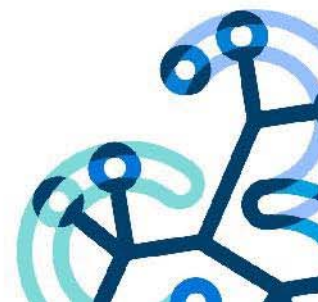
日志异常检测算法

- 对故障时段的日志做聚类形成日志模式，并与正常时段的日志模式做对比，发现异常日志行为 定位故障原因
- 价值：
 - 出现故障后迅速定位原因，减少 MTTR
 - 弥补现有监控规则覆盖不到的故障情况
- 算法特点：
 - 无需对日志做解析
 - 针对监控规则没覆盖的场景，价值大，准确率高



日志异常检测算法处理步骤

- 初步聚类：层次聚类（专利）
- 频繁词权重计算
- 词与词的依赖权重计算
- 多次迭代
- 聚类内模式提取
- 性能：每秒 3 万 + 条日志

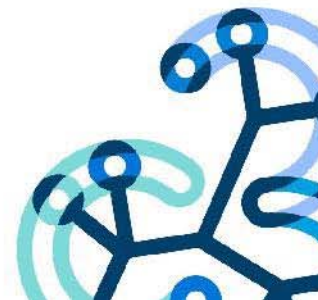


故障消息聚类（一）

故障聚类：将同一个故障引发的告警聚类为故障模式

一、故障消息相似度，特征包括：

- 时间：窗口（白天窗口，夜间窗口、节假日窗口）
- 描述：停止词、行业同义词（互联网、金融）、实体、动作
- 主机/服务：拓扑距离、数据链路距离
- 级别：
- Agent：



故障消息聚类（二）

故障聚类：将同一个故障引发的告警聚类为故障模式

二、故障消息中关键词的关联性：

- disk 80% usage
- network failure

三、历史数据学习

- 故障消息之间的关联性、主机之间的关联性
- 分析故障 MTTR，调整聚类时间窗口



故障消息聚类评价指标

任意两个告警 e_1 和 e_2 在分配给情境时会发生以下四种情况:

- TP (True Positive) 分配: 属于同一个真实情境的告警 e_1 和 e_2 , 被分配到同一个推断情境中。✓
- TN (True Negative) 分配: 不属于同一个真实情境的告警 e_1 和 e_2 , 被分配到不同的推断情境中。✓
- FP (False Positive) 分配: 不属于同一个真实情境的告警 e_1 和 e_2 , 被分配到同一个推断情境中。✗
- FN (False Negative) 分配: 属于同一个真实情境的告警 e_1 和 e_2 , 被分配到不同的推断情境中。✗



故障根因定位

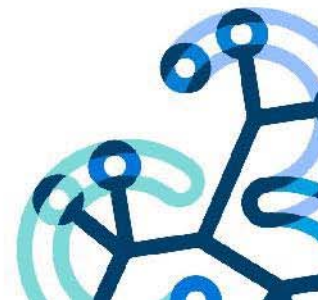
根因分析假设：历史故障越频繁，异常程度越高的数据越可能是根因，难点是如何计算数据的异常程度，即信息熵

- 故障消息的信息熵：消息内容、来源主机、时间
- 日志的信息熵：错误日志的数量、日志模式对比（专利）
- 变更：一上线就挂，信息熵高
- 时序指标：预测、异常检测（无监督 / 有监督）



客户案例一

- 某业务因 Redis 内存超限发生故障
- 2月16日 Redis 故障回放
- 08:58 大屏人员看到告警 “xxxx- 总体 API 异常”
- 09:09 大屏人员联系业务 RD
- 09:25 业务 RD 通知系统运维，初步确定为 Redis 问题
- 09:35 系统运维确认问题为 Redis 节点内存使用超限
- 09:40 解决问题，业务逐步恢复正常



客户案例一：效果 - 告警聚类

[返回上一级](#) | 故障模式: 977

创建时间: 2019-02-16 08:58:04 | 故障模式状态: 待接手 | 关联故障数: 2 | 故障模式持续时间: 51分21秒 | 关联日志数: 0

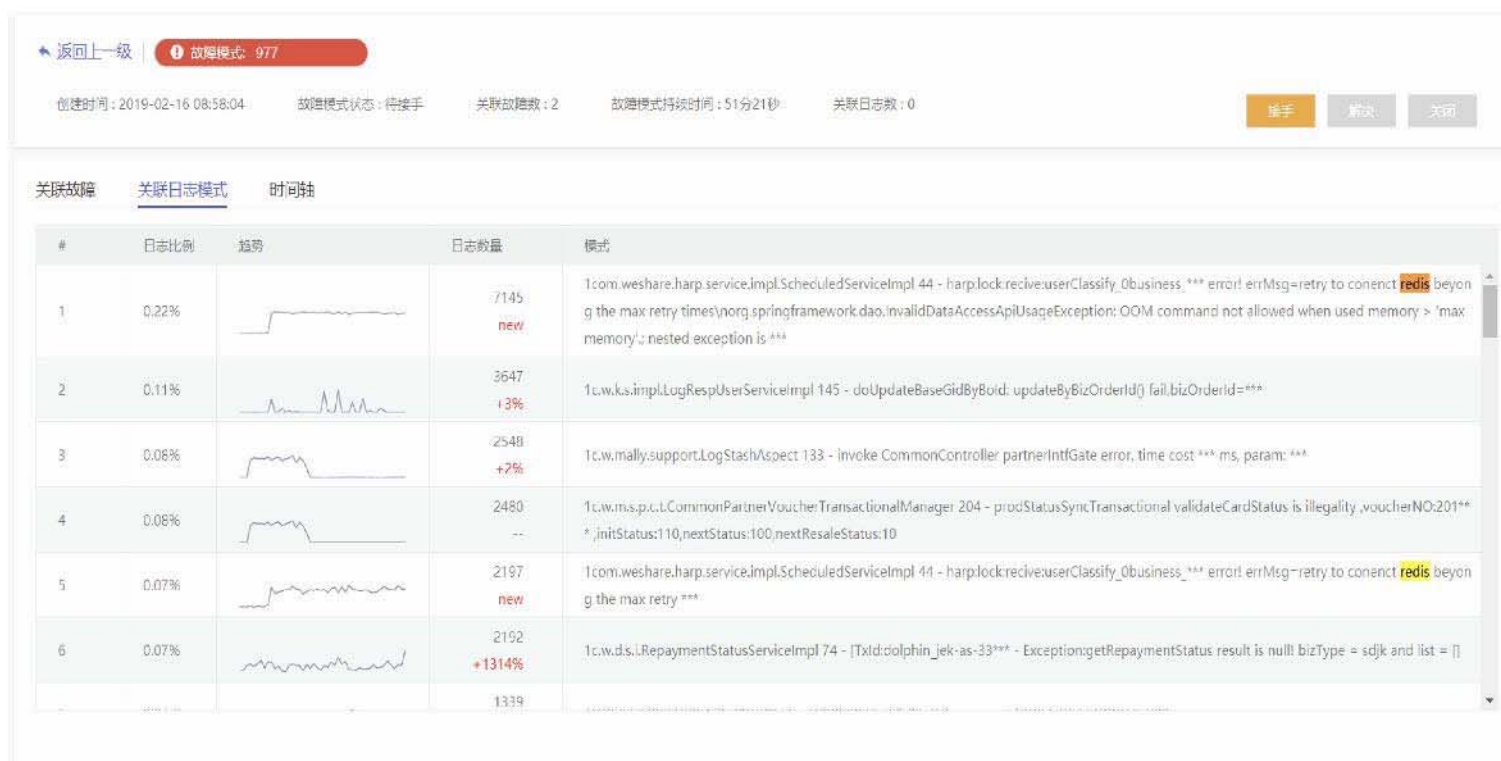
[接手](#) [解决](#) [关闭](#)

[关联故障](#) | [关联日志模式](#) | [时间轴](#)

	故障等级	ID	根因评分	故障内容	发生踪	故障类型	产生时间	合并故障日志数
⊖	重要 42		0.878	总体API异常 -- 10分钟内总体API异常的数...		闪电借款2.0 -- AP...	2019-02-16 08:58:04	2
○	重要 44		0.316	总体API异常 -- 10分钟内总体API异常的数...		现金贷通 -- API后...	2019-02-16 09:49:25	1



客户案例一：根因排查可节省 15~20 分钟



客户案例二

- 某业务因机器负载过高出现失败
- 15:51 业务监控报出总体 API 异常
- 16:41 业务方反馈部分业务访问 Redis 异常
- 16:50 运维确认一台 Haproxy 节点压力过高
- 之后运维开始修复操作



客户案例二：结果展示 - 告警聚类

[返回上一级](#)
故障模式: 1300

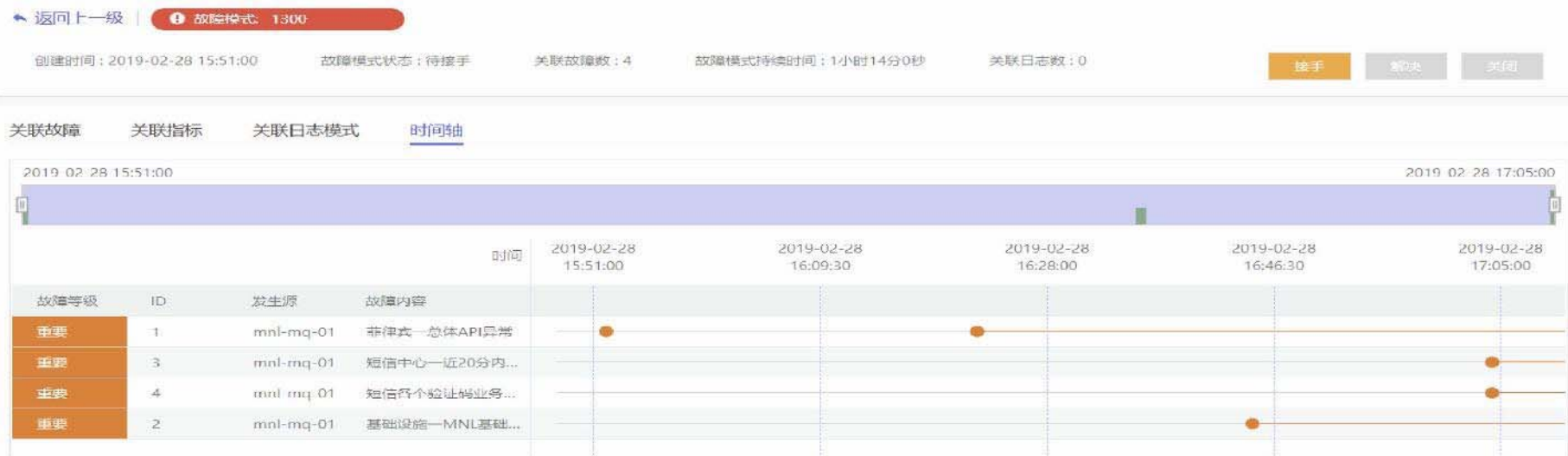
创建时间: 2019-02-28 15:51:00
故障模式状态: 待接手
关联故障数: 4
故障模式持续时间: 1小时14分0秒
关联日志数: 0
接手
解决
关闭

[关联故障](#)
[关联指标](#)
[关联日志模式](#)
[时间轴](#)

	故障等级	ID	根因评分	故障内容	发生源	故障类型	产生时间	合并故障消息数
🔍	🔴 重要 1		0.865	菲律宾-总体API异常	mnl-mq-01	菲律宾	2019-02-28 15:51:00	2
🔍	🔴 重要 3		0.325	短信中心一近20分钟内短信量对比	mnl-mq-01	菲律宾	2019-02-28 17:05:00	1
🔍	🔴 重要 4		0.325	短信各个验证码业务端条数	mnl-mq-01	菲律宾	2019-02-28 17:05:00	1
🔍	🔴 重要 2		0.313	基础设施-MNL基础监控	mnl-mq-01	菲律宾	2019-02-28 16:45:00	1



客户案例二：结果展示 - 时间轴



客户案例二：实时定位异常指标

[返回上一级](#)

故障模式: 1300

创建时间: 2019-02-28 15:51:00

故障模式状态: 待接手

关联故障数: 4

故障模式持续时间: 1小时14分0秒

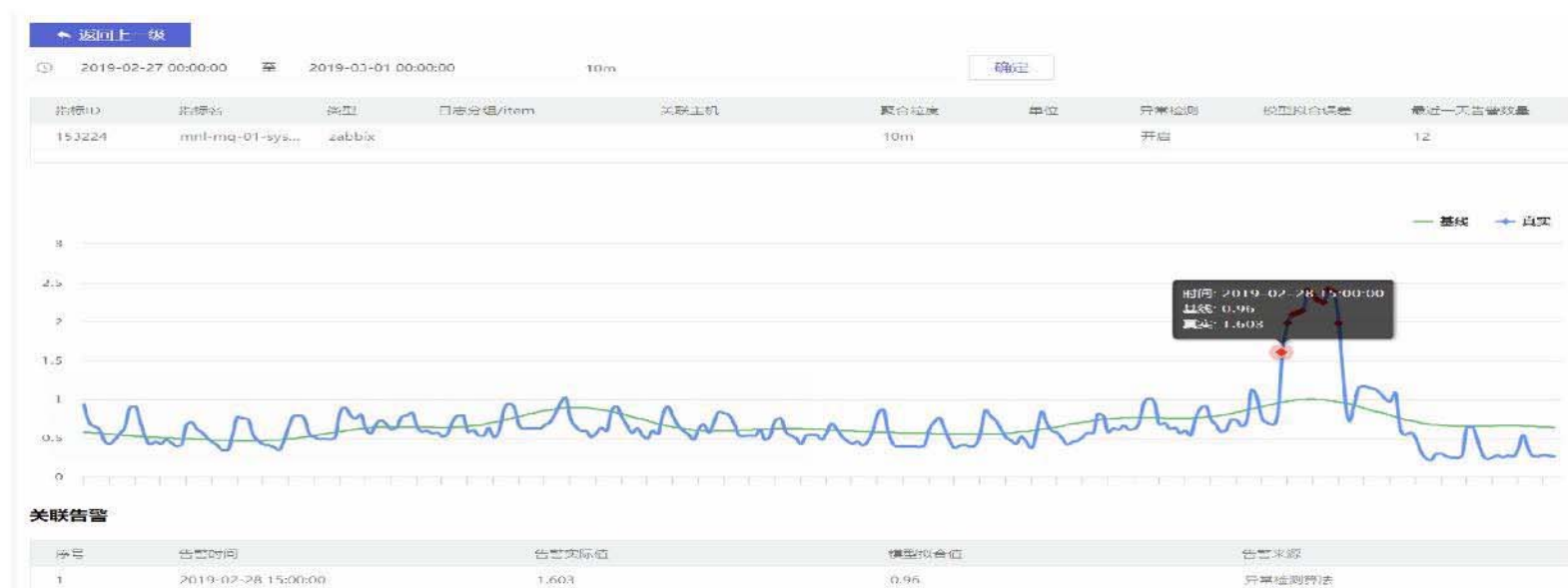
关联日志数: 0

[接手](#)
[解决](#)
[关闭](#)
[关联故障](#)
[关联指标](#)
[关联日志模式](#)
[时间轴](#)

指标名	类型	聚合粒度	单...	数据点个数	实际值	基线值	同比	状态
mnl-mq-01-system.cpu.load[percpu,avg5]	zabbix			10m		20	1.64	0.96	57.9%	异常
mnl-mq-01-system.cpu.load[percpu,avg15]	zabbix			10m		20	1.62	0.96	57.3%	异常
mnl-mq-01-system.cpu.load[percpu,avg1]	zabbix			10m		20	1.6	0.94	58.4%	异常
mnl-mq-01b-system.cpu.load[percpu,avg15]	zabbix			10m		20	0.55	0.57	-2.6%	正常
mnl-mq-01b-system.cpu.load[percpu,avg5]	zabbix			10m		20	0.53	0.57	-4.5%	正常
mnl-mq-01b-system.cpu.load[percpu,avg1]	zabbix			10m		20	0.45	0.49	-6.8%	正常
mnl-mq-01c-system.cpu.load[percpu,avg1]	zabbix			10m		20	0.14	0.16	-5.8%	正常



客户案例二：异常检测算法： 可提前大屏监控 51 分钟，提前业务侧 101 分钟发出告警



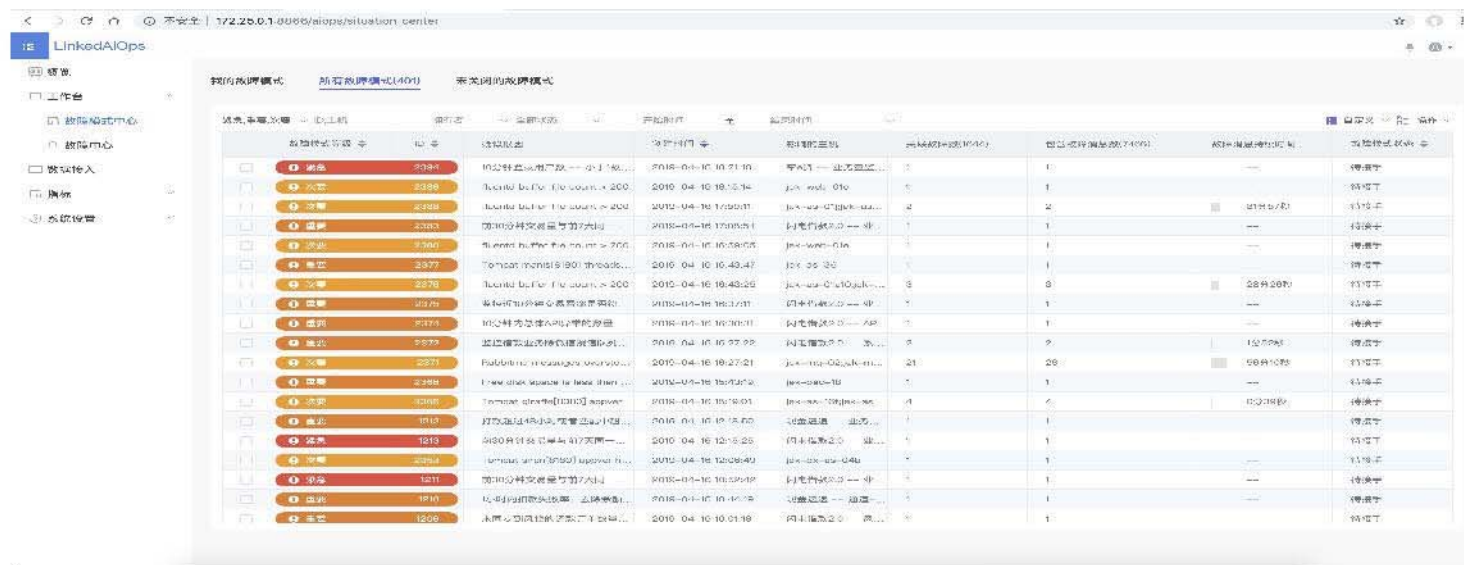
- 试运行客户使用场景
使用触发条件：
 - 客户收到原监控系统发出的告警，需要寻找根因或确定告警影响范围时，登陆产品
- 产品使用方法：
 - 根据收到告警的时间、关键字、主机信息，在情境中心的最前面找到描述相同的情境，点击查看分析结果详情
- 分析结果使用方法：
 - “关联指标”列表的最前面会显示异常指标和所在主机，点击查看趋势图可做人工确认
 - 出现业务告警时，“关联日志”中显示为 “new” 和数量突增的日志需要引起关注，可通过日志模式文本判断根因和故障影响范围



- 场景选择方法：
 - 在试运行汇报现场，客户现场使用产品，找到最近收到的告警在产品中的位置，查看产品分析结果
- 最近告警和故障内容：
 - 业务做了消息推送，导致短时间内客户大量访问，触发系统层中间件和消息队列负载过高的告警
- 客户对产品输出的期望：
 - 希望产品能将本次故障涉及的很多条告警聚类到一起
 - 希望产品能发现故障关联的主机有 CPU 异常升高，且故障之前一天没有异常



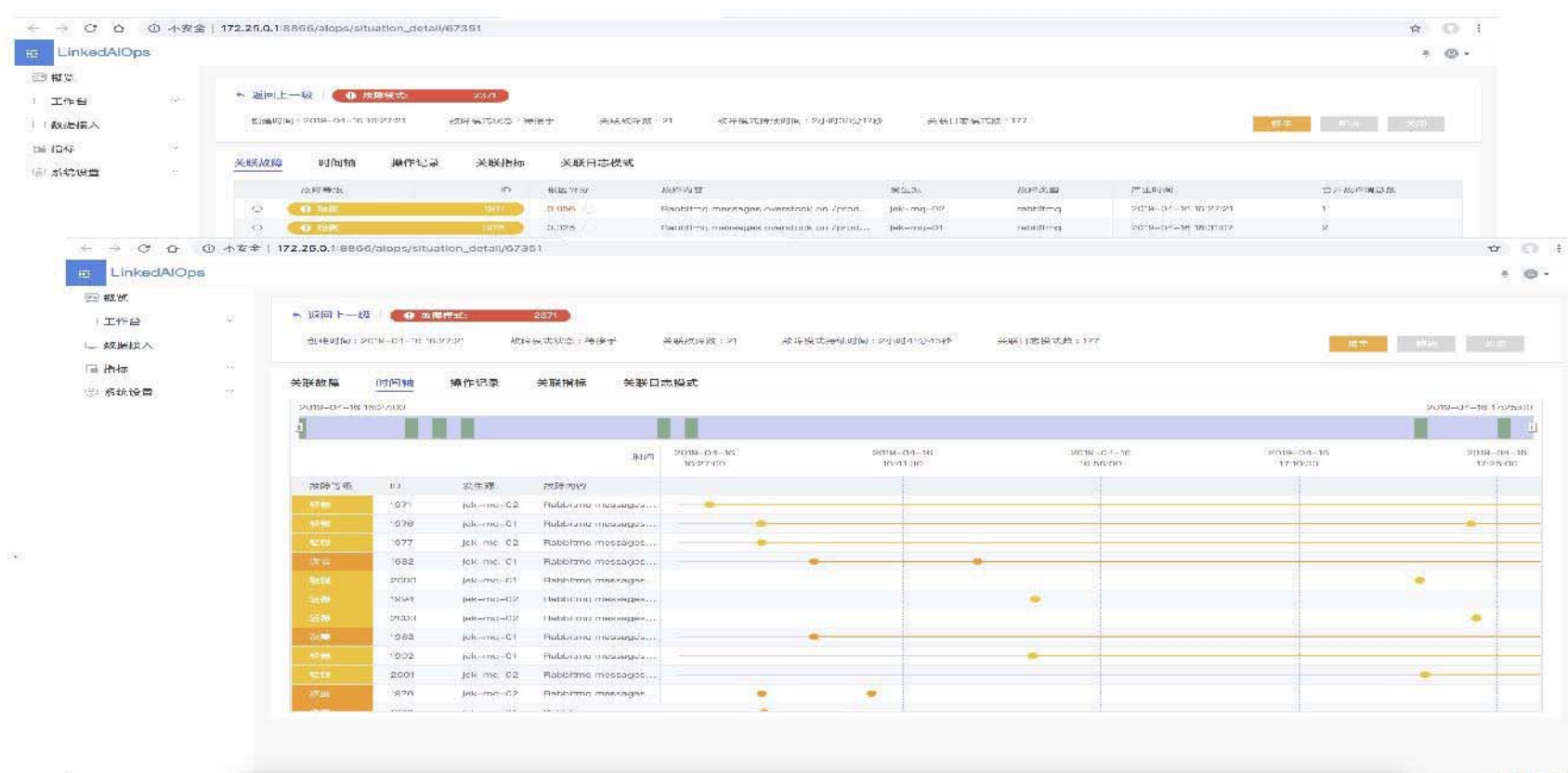
试运行场景一：客户从该页的 **2371** 情境找到与告警短信对应的 Rabbitmq 告警，共花费了十秒钟左右



告警名称/来源	ID/序号	详细状态	发生时间	影响的资源	关联的告警	告警级别	告警状态
告警	2384	10分钟业务量下降...	2018-04-10 10:21:10	应用...	告警...	1	待处理
告警	2388	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2388	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	2	待处理
告警	2383	10分钟业务量下降...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2380	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2377	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2378	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2379	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2375	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2373	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2371	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	20	待处理
告警	2368	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	2366	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	1813	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	1811	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	1810	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理
告警	1209	Redis buffer for...	2018-04-10 10:15:14	应用...	告警...	1	待处理



试运行场景一：关联告警，合并效果客户认为符合预期



试运行故障一

- 故障集中在 mq-01, mq-02 两台机器, 模型 mq-01 cpu 出现异常, 符合预期, 客户希望看指标是否人工确认是否异常
- 异常出现的指标与推送指标有十分左右延迟, 符合预期, 说明成功召回异常; 且推送之前模型未出现异常, 说明模型没有产生

