ArchSummit
全球架构师峰会 2017

Alibaba Group 阿里巴巴集团 GOC Global Operations Center 主办方 Geekbang 极客邦科技 InfoQ

阿里巴巴故障治理领域的智能运维实践

王肇刚(梓弋)

阿里巴巴集团 GOC事业部 高级技术专家

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

主办方 Geekbang. 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

SPEAKER
INTRODUCE

王肇刚（花名：梓弋）

阿里巴巴集团 GOC事业部 高级技术专家

负责阿里巴巴集团业务指标监控、业务故障管理工作。在时间序列异常检测、业务故障定位及影响面分析、运维数据仓库和其它相关的智能运维领域有丰富的技术经验积累和成果产出。

在加入阿里巴巴之前，作为百度智能运维团队的架构师及核心项目负责人，主导了服务于百度商业广告系统异常发现和故障定位的智能运维产品的设计和研发，并主导了百度运维数据仓库及百度智能运维平台的设计和研发工作。

TABLE OF

CONTENTS 大纲

- 阿里巴巴故障治理业务流程及挑战
- 引入智能运维的效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

双11峰值背后的挑战巨大

每秒
32.5万笔
订单

每秒
25.6万笔
支付

业务线的多样和复杂给服务稳定性带来挑战

- 业务数量及规模不断增大
- 业务形态差异较大
- 业务关联复杂

线上故障需要统一的治理机制

淘宝 Taobao

天猫 TMALL.COM

AliExpress Smarter Shopping, Better Living!

蚂蚁金服 ANT FINANCIAL

支付宝 ALIPAY

MORE THAN JUST CLOUD | Alibaba Cloud

业务故障统一发现

跨BU故障协同处理

故障的影响面和根因
需要统一收口和推送

故障快速恢复需要
统一的机制

阿里巴巴全局故障治理流程

故障发现 → 故障定级 → 故障通告 → 故障辅助定位 → 处理决策 → 故障快速恢复 → 故障复盘 → 故障演练

业务流程

传统监控系统误报漏报较多

监控维护成本较大

故障等级定义差异较大

判断条件繁多

千万级别的运维事件，哪些与业务故障相关？

跨BU的应用依赖复杂，如何梳理追溯

快速恢复场景稍纵即逝，如何实时决策触发切换？

业务痛点

全球运行指挥中心

TABLE OF

CONTENTS 大纲

- 阿里巴巴全局故障治理业务流程和挑战
- 引入智能运维的场景和效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

引入智能运维的场景和效果

TABLE OF

CONTENTS 大纲

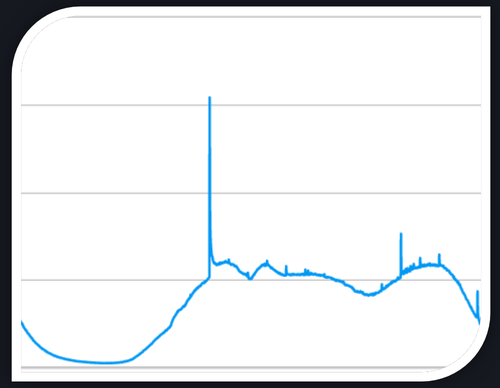
- 阿里巴巴全局故障治理流程和业务痛点
- 故障治理领域引入智能运维的效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

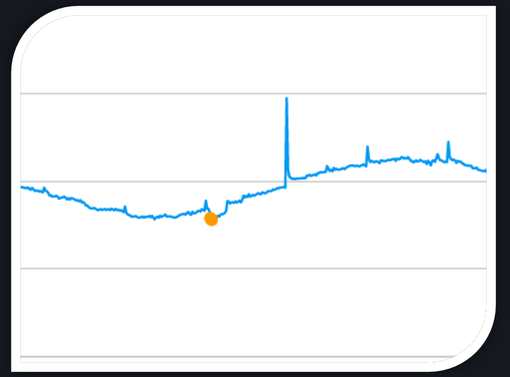
实战案例：业务异常发现

异常发现的业务背景

“淘宝交易量下跌%X是Pn故障”

… …



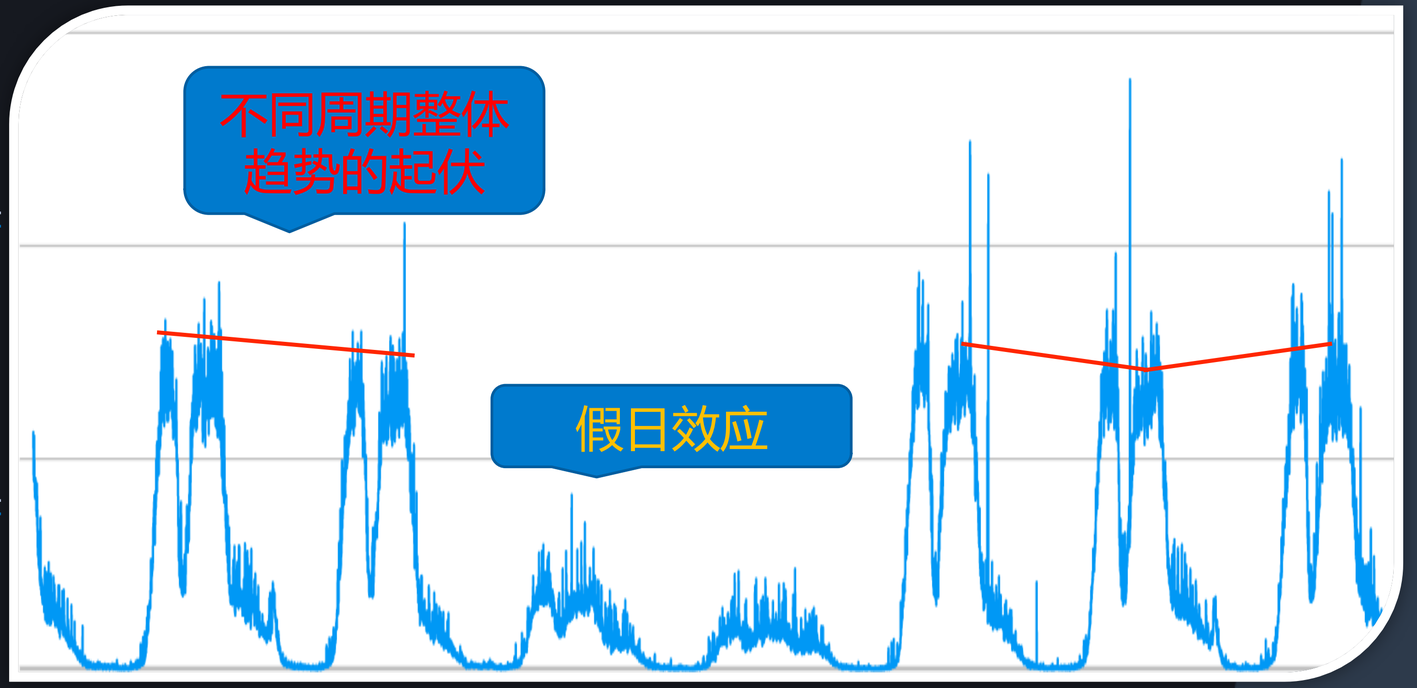


“[Pn][淘宝]淘宝交易创建下跌X%”

故障等级定义 → 业务指标监控项（时间序列） → 异常点 → 故障通告

异常发现的业务痛点—如何确定基线

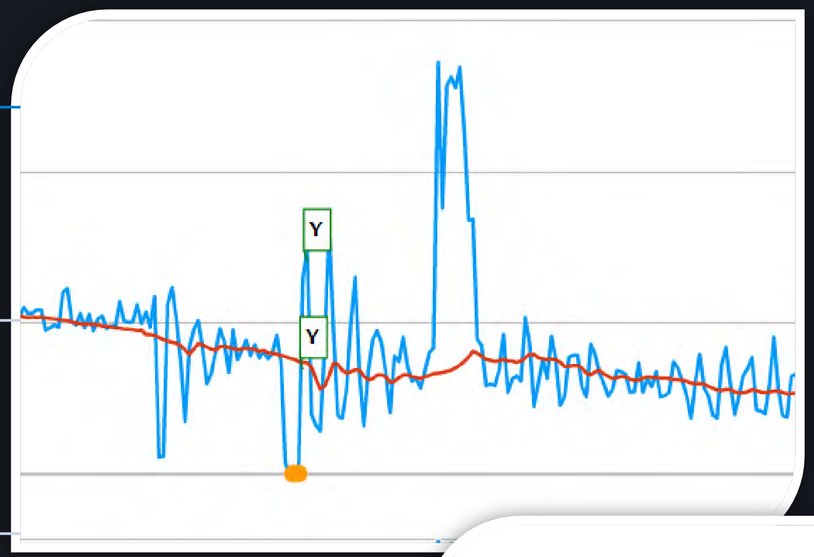
问题	“下跌15%是和什么相比下跌？”
分段静态阈值	无法应对业务局部趋势变化
同环比/过去N周分段均值	无法应用业务整体起伏趋势

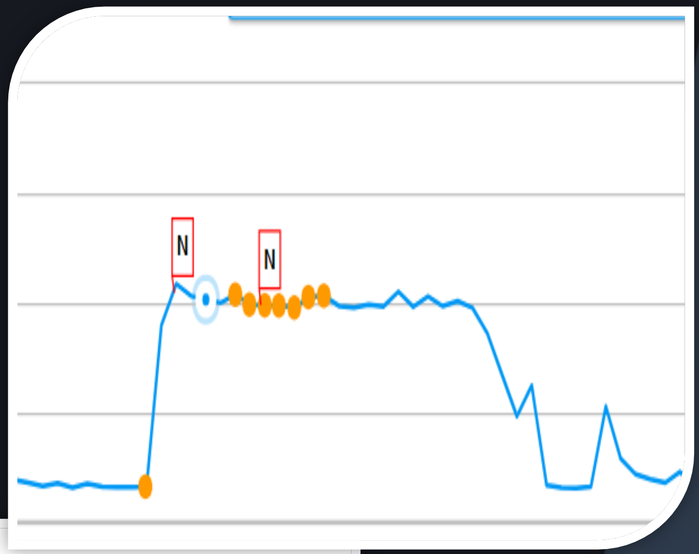


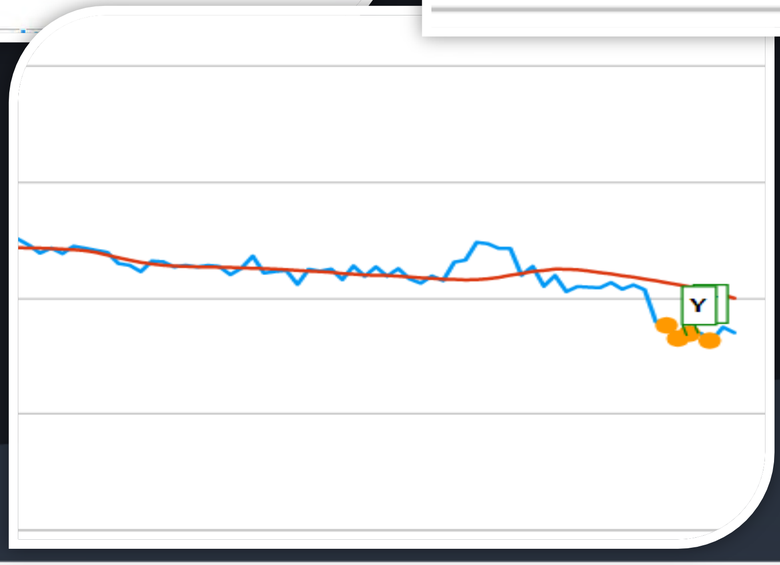
异常发现的业务痛点—如何判定异常

业务异常的判定尺度

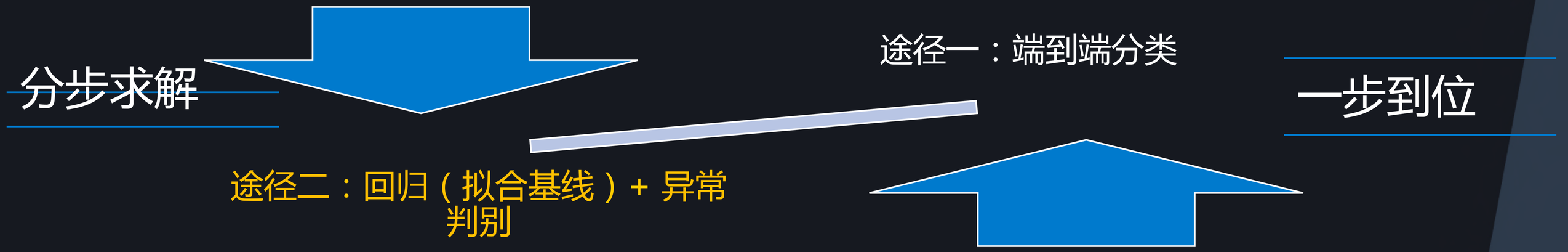
- 与曲线本身波动程度相关
- 与曲线宏观业务量相关
- 与时间点相关
- 与业务特性相关







时间序列异常检测的方案选择



回归	各类机器学习模型
	训练样本充足
异常判别	依赖标注
	标准不统一

基于机器学习/深度学习

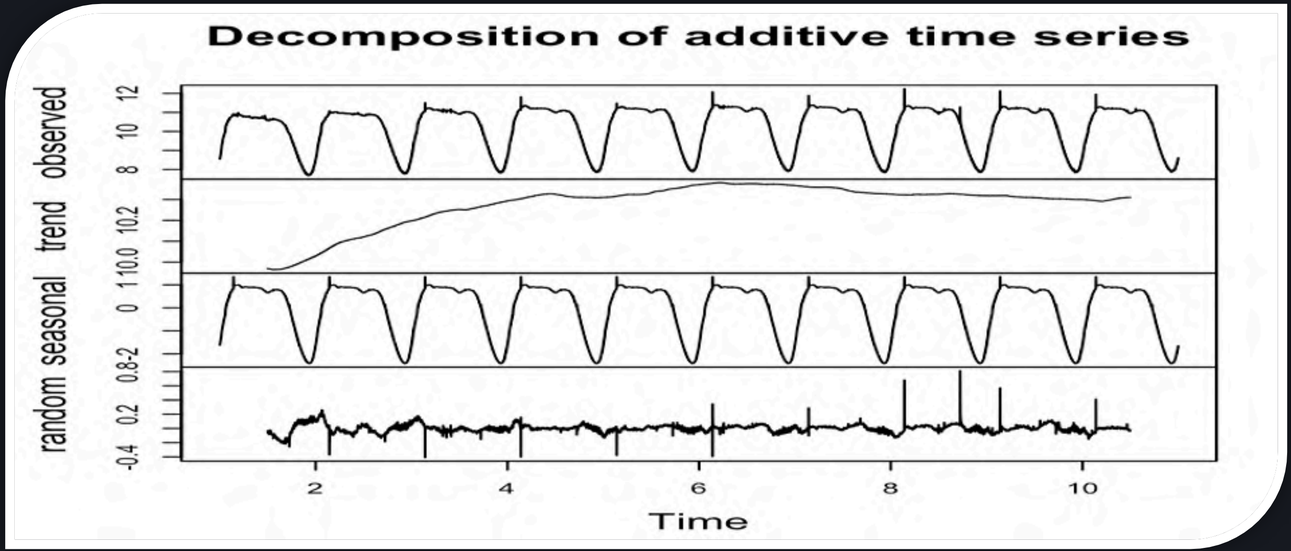
基于时间序列分解

回归	各类时序分解算法
	方法可解释性强
异常判别	简单策略
	复合方法

时间序列在线预测：拟合基线

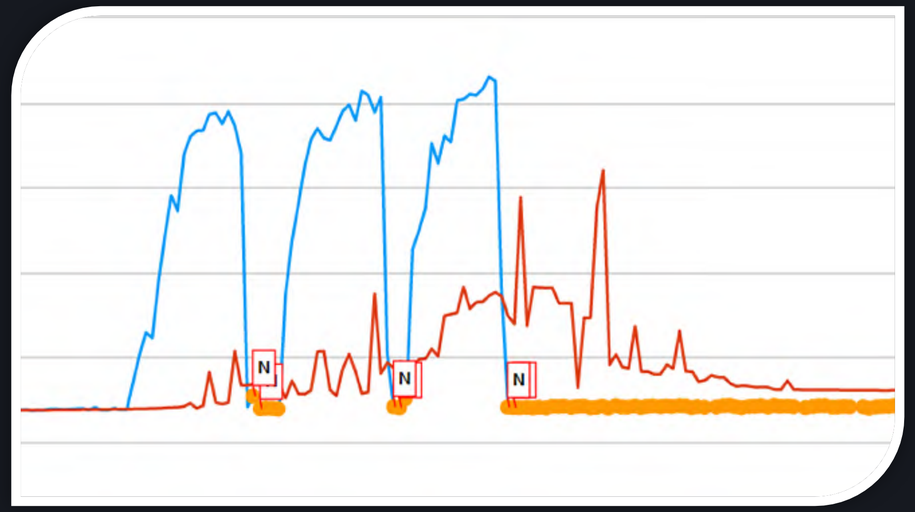
可选方案
- 分段历史平均
- ARIMA
- Holt-Winters
- STL

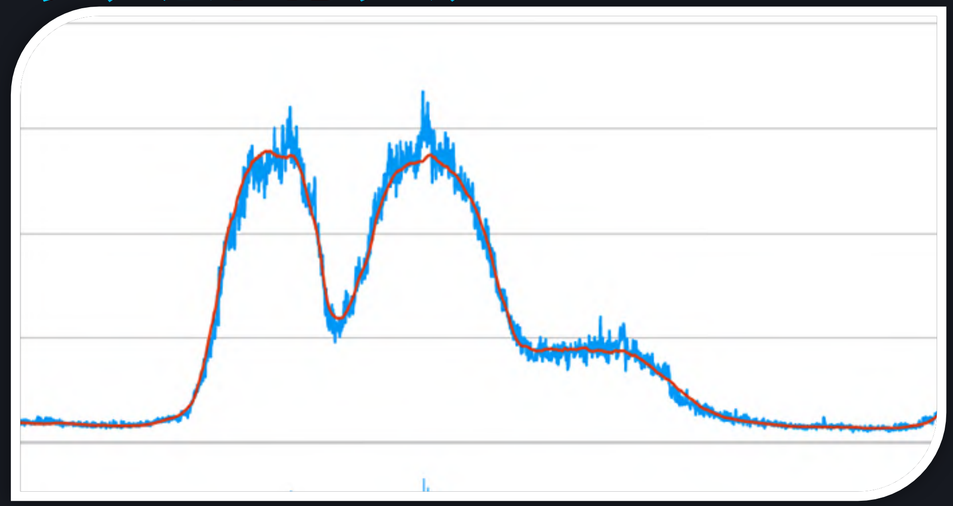
质量控制
- 量化评估
- 基线质量分



KPI NAME	Holt-Winters	历史数据平滑	ARIMA	STL
淘宝交易创建	0.27	0	1.0	1.7
搜索|广告-主搜店铺	0.27	0.38	0.3	1.68
手淘直充成功量	1.38	0.56	1.2	0.3
聚划算交易与创建	0.57	0.39	0.57	2.72

拟合基线的关键步骤：数据预处理





丢点补全

日期类型划分

预测“未来”

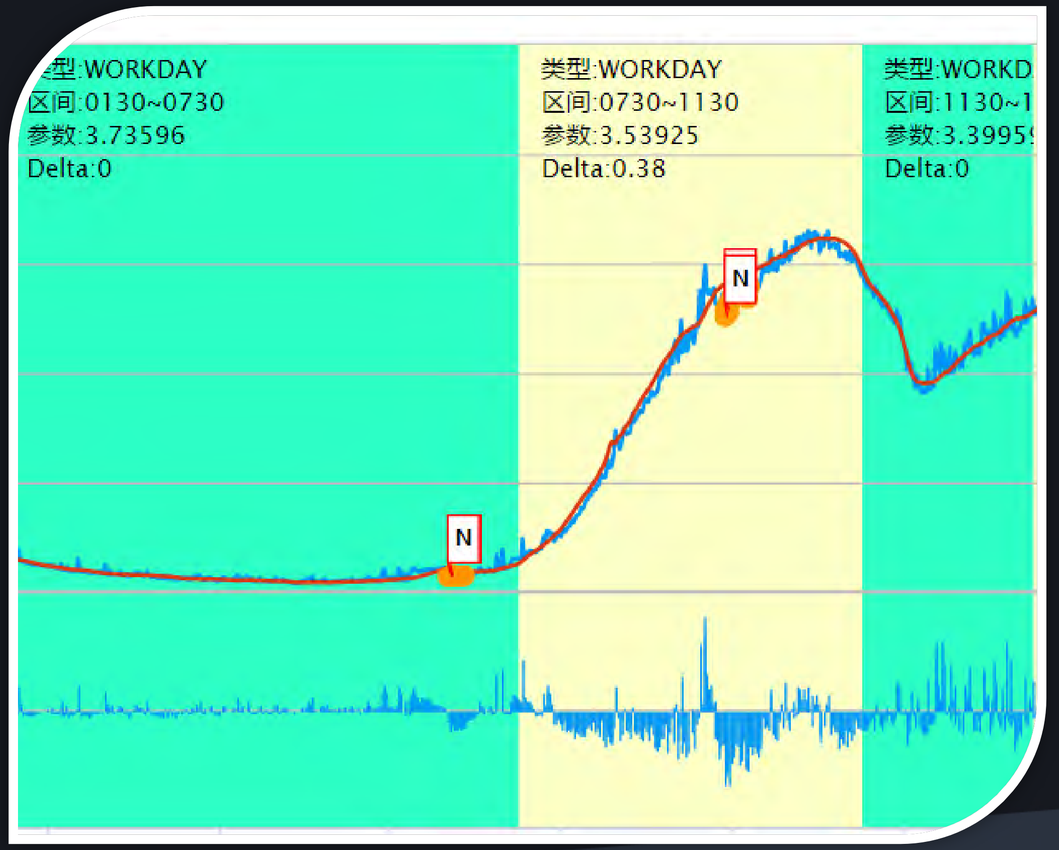
局部趋势反馈

异常判定—X倍-Sigma

时间片切分	根据残差分布进行聚合

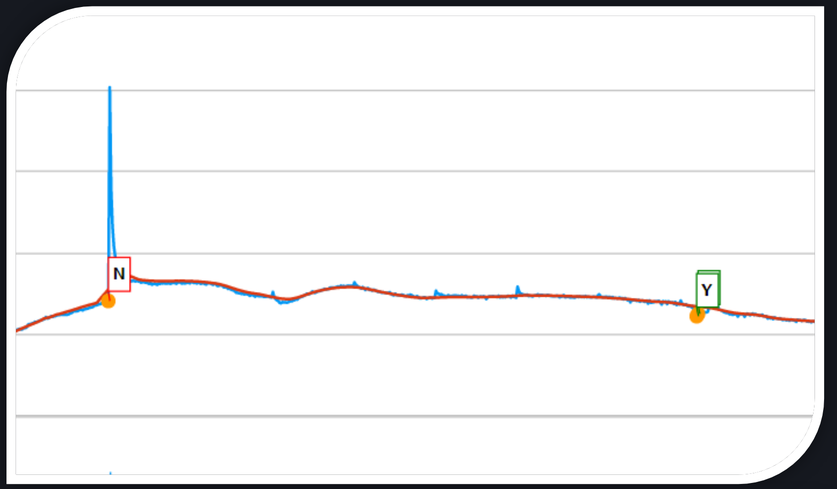
每个时间片的X=N+Delta

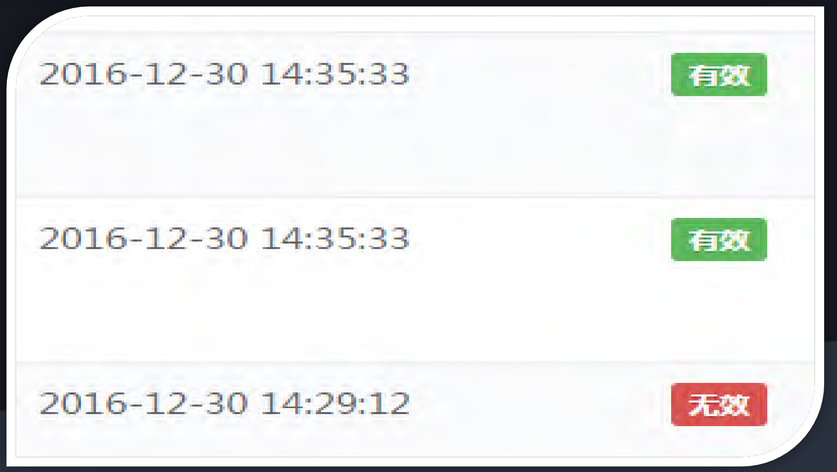
N	与残差相关
Delta	与用户反馈相关



异常判定—用户标注反馈

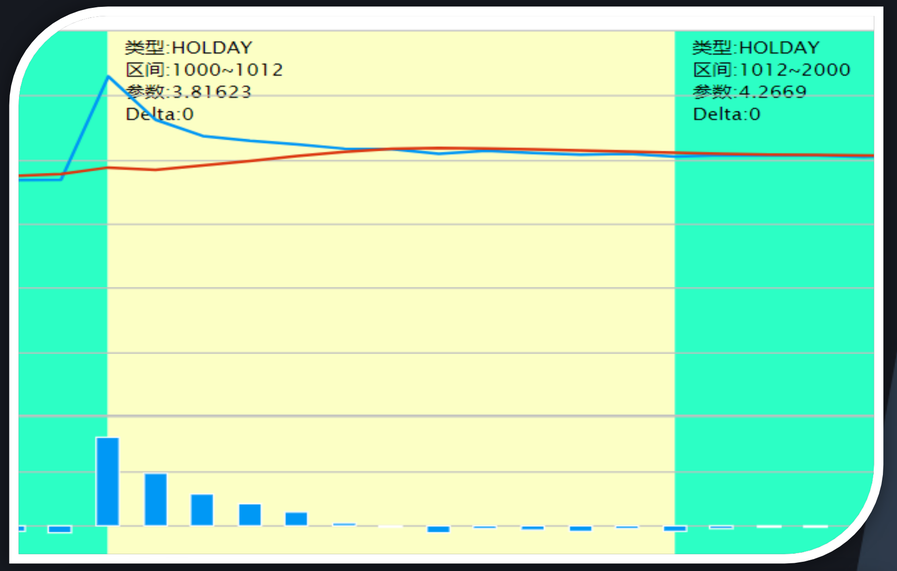
关于标注	为运营而生的打标数据
	标注数据质量较差
根据标注调整Delta	容忍标准误差
	防止参数抖动

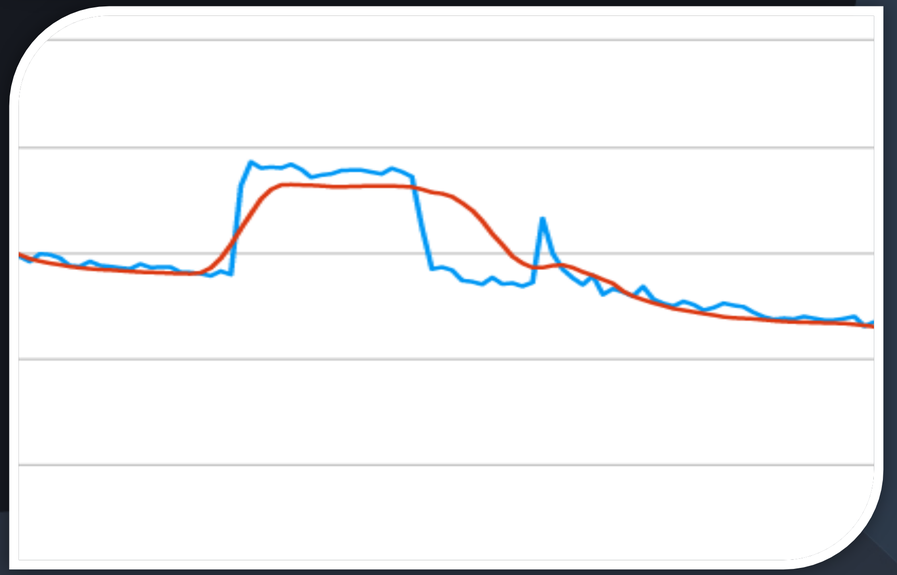




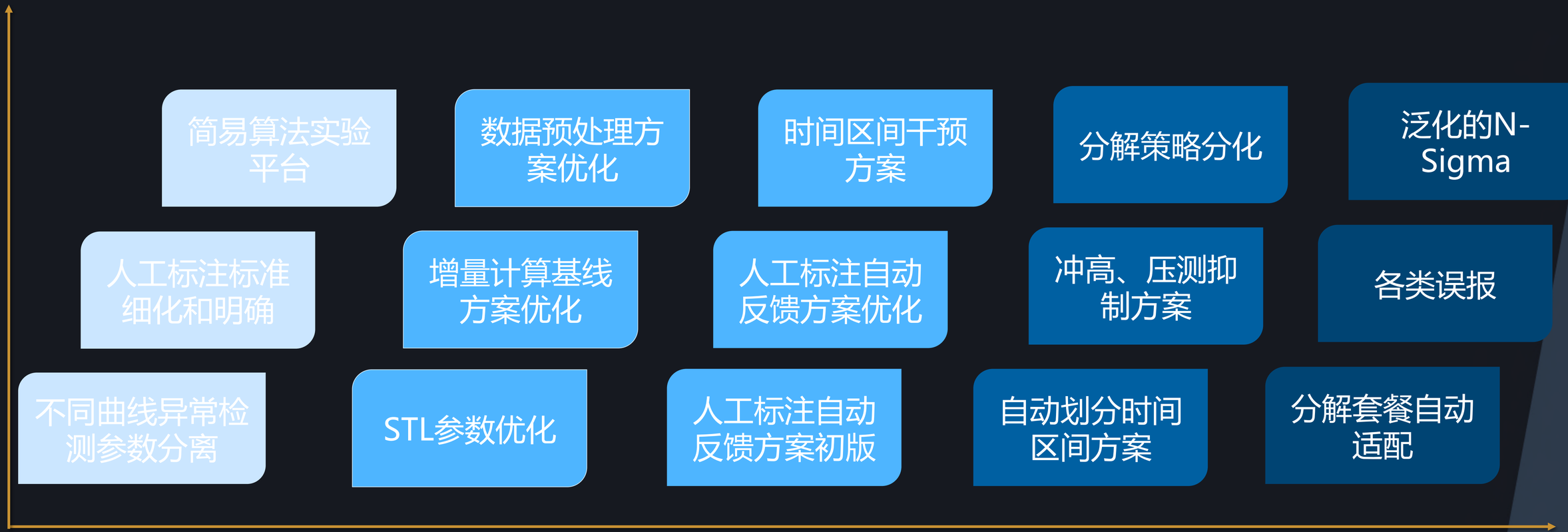
异常判定– 误报抑制

冲高回落时的误报抑制	基线预处理	
	分段策略	
基线不准时的误报抑制	基线质量在线检查	压测状态
		基线长期偏离状态
	投票策略	跳变检测
		局部特征





异常检测算法及工程演进历程

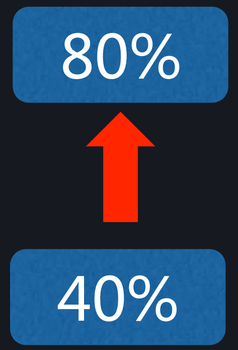


异常检测系统工程架构

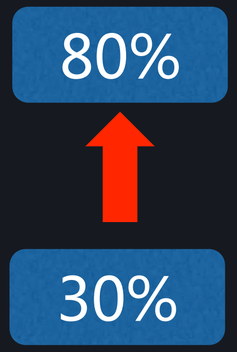

展示输出
算法逻辑
前端WEB展示
API
基线输出
异常检测输出
运行模式
生产者
消费者
功能
拟合基线
异常检测
数据
监控数据源适配
监控项管理
趋势数据存储
报警和用户标注
基础组件
分布式定位调度
TSDB
Redis
算法参数管理
调度任务
基线计算
异常检测
冲高回落
去除压测
对外数据
推送
打标反馈
检测参数
初始化
自动分段
自动分类
清除缓存

异常发现业务效果

故障发现准确率



故障发现召回率



每周因误报而花费的流程操作时间

29小时

但，异常发现还不等于故障定级

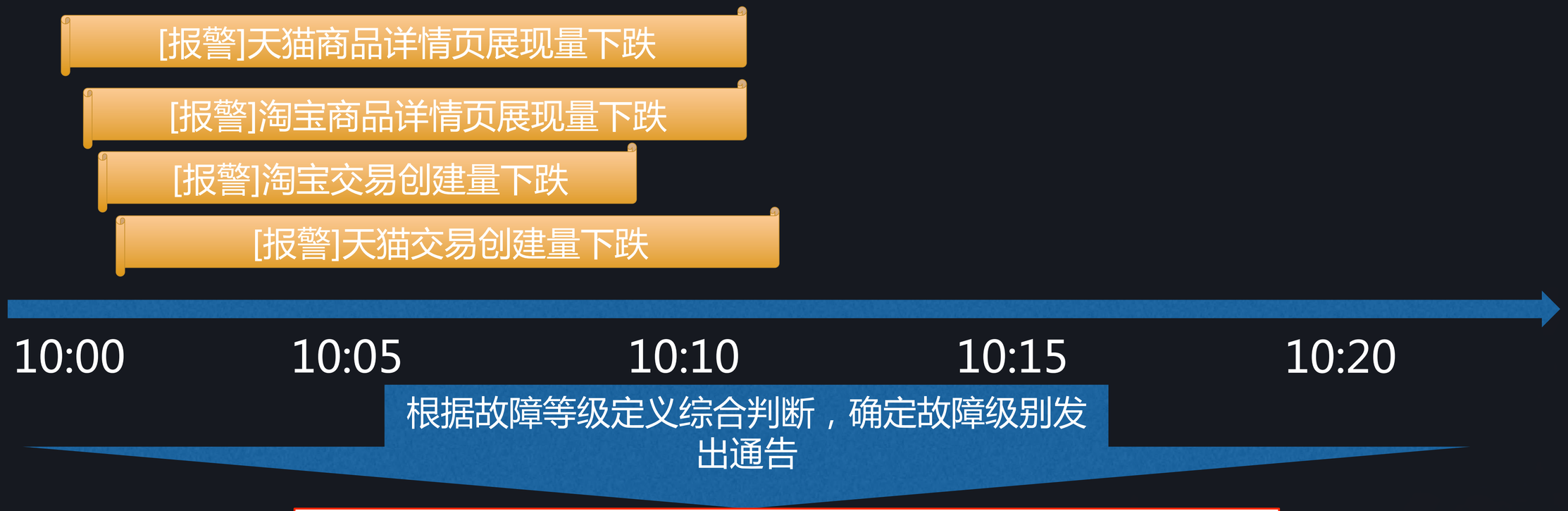
TABLE OF

CONTENTS 大纲

- 阿里巴巴全局故障治理流程和业务痛点
- 故障治理领域引入智能运维的效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

案例实战：故障智能定级

业务流程



[故障通告][XX事业部][P2]于10:01，淘宝、天猫交易创建下跌X%，相关人员正在处理中。[GOC]

案例实战：故障智能定级

业务痛点

“淘宝交易量下跌%X是Pn故障”

- 输入监控项维度复杂
- 判断条件复杂
- 报警聚合收敛依赖经验
- 维护成本高

“淘宝交易量下跌%x且持续y分钟，同时至少z各个分机房持续下跌持续m分钟，是Pn故障。同时，如果影响多个BU，则以故障等级严重的定义为准。”

- 理想中的故障等级定义
- 现实中的故障等级定义

解法：报警智能收敛和定级

报警纵向收敛 → 报警横向收敛 → 故障级别确定 → 故障通告/更新

同业务报警收敛

按时间窗口收敛

跨业务报警收敛

依赖业务拓扑收敛

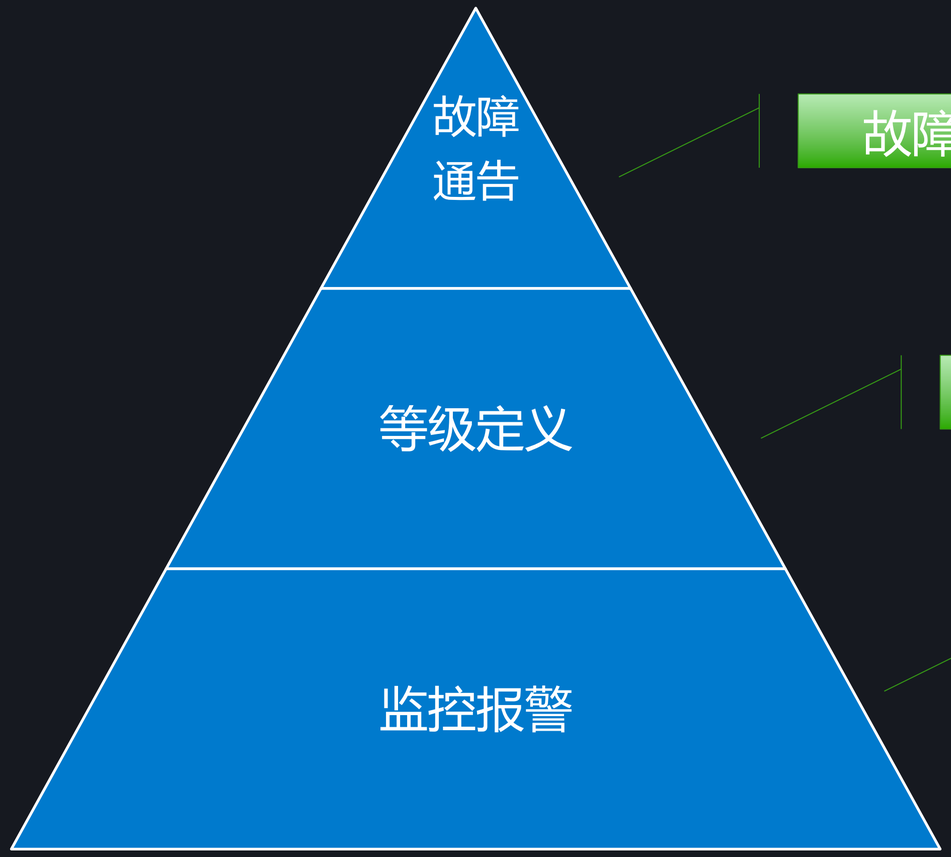
故障定级

- 故障等级定义结构化
- 规则引擎定级
- 从重定级

通告或更新

- 之前是否有未结束的故障
- 当前故障与否与历史故障同属一个故障

挑战之一：数据结构化



故障通告结构化

- 故障影响面结构化
- 故障根因结构化
- 故障处理行为结构化

等级定义结构化

- 数据来源标准化
- 定级标准统一化
- 联合判断规则化

监控数据标准化

- 监控对象标准化
- 监控项维度层次梳理
- 异常程度输出标准化

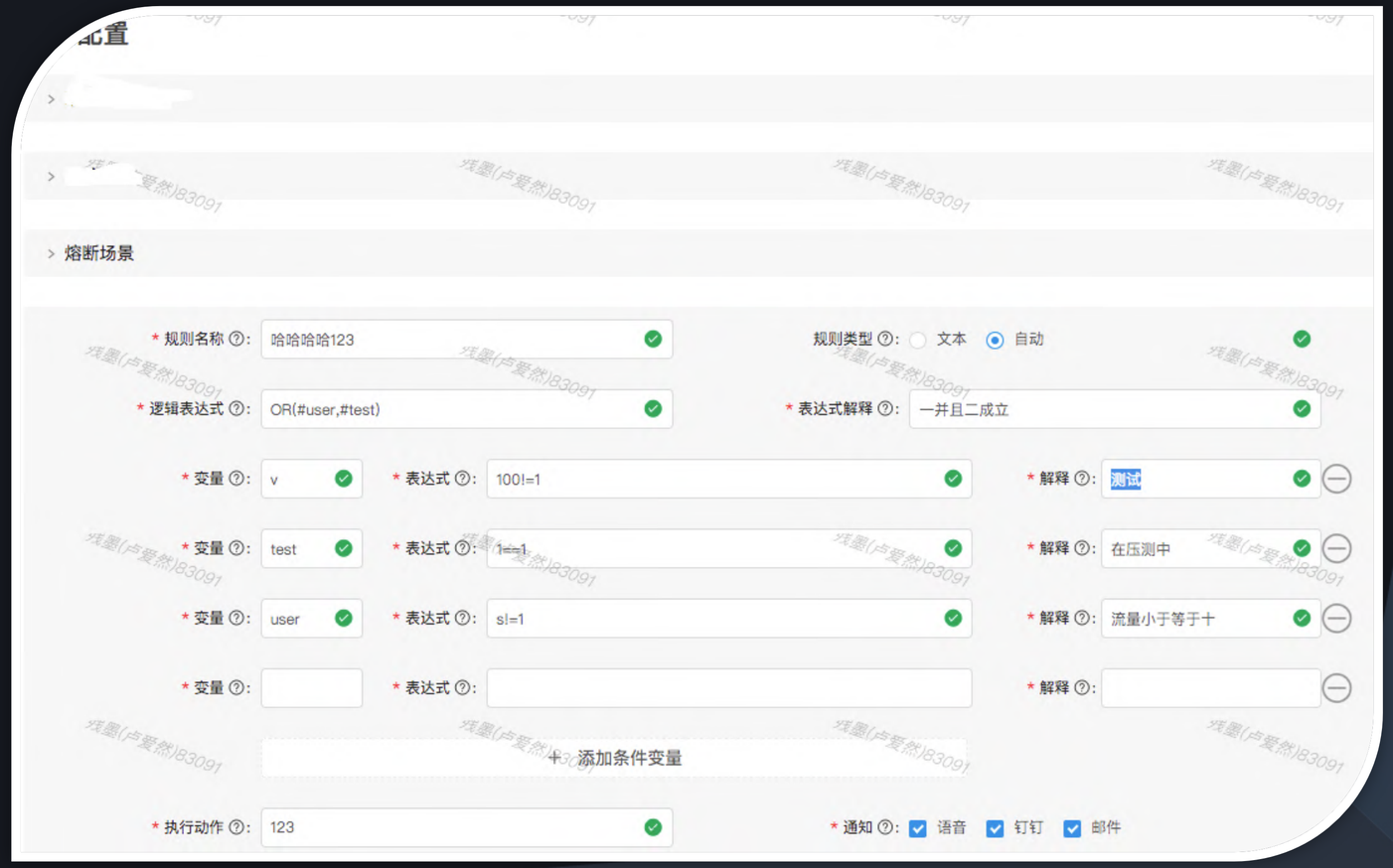
挑战之二：规则引擎

格式化判断输入

格式化判断动作

格式化判断输出

多层逻辑表达式组合



故障智能定级效果

首条故障通告SLA

5分钟 → 1分钟

故障快速切换决策时间

？分钟 → 1分钟

异常发现，故障定级，我们是否还能做得更多？

TABLE OF

CONTENTS 大纲

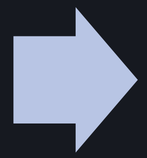
- 阿里巴巴全局故障治理流程和业务痛点
- 故障治理领域引入智能运维的效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

实战案例：故障智能分析

故障自动分析及定位的难点

故障分析定位的范围及边界的确定

故障分析定位的信息收集



故障分析定位的判断和决策逻辑

故障分析定位的范围和边界

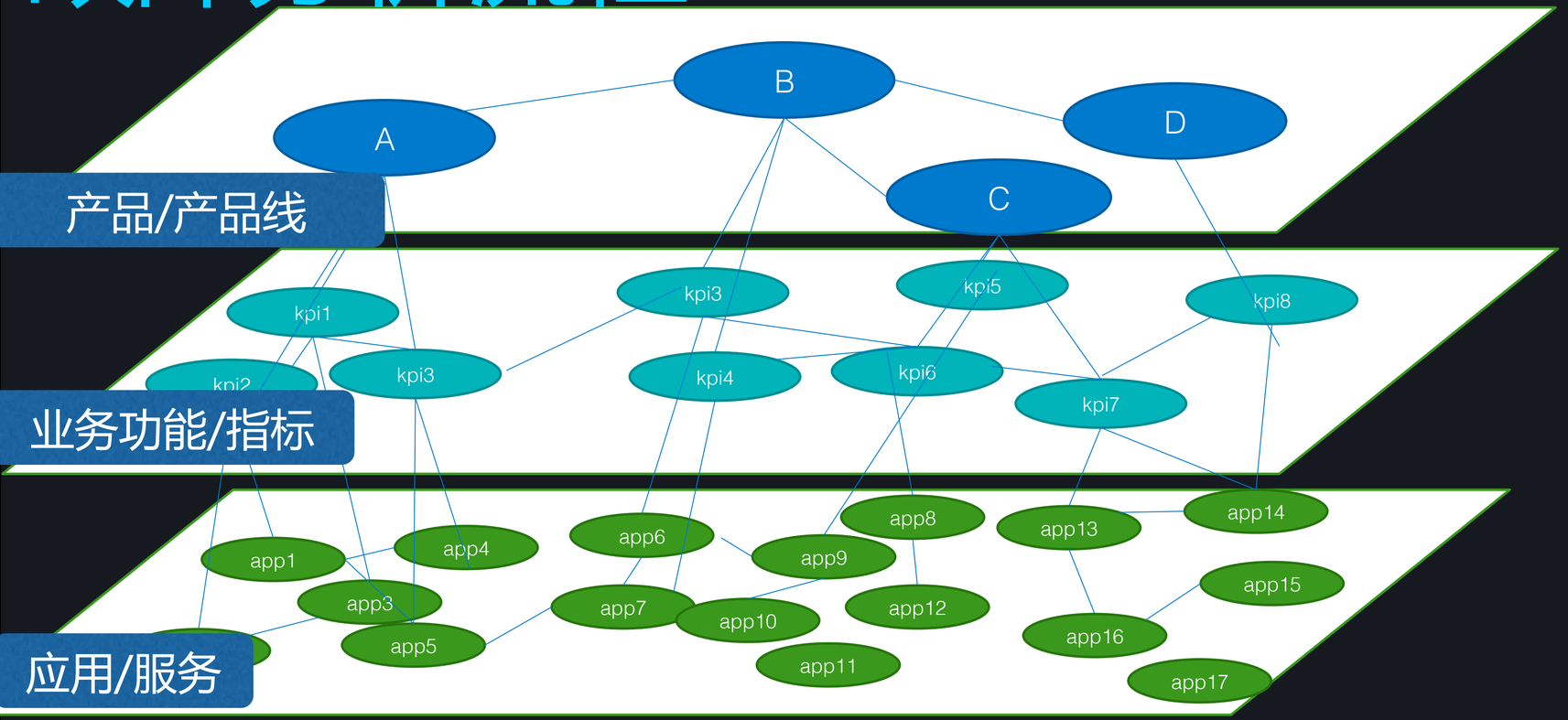
触发条件 业务指标异常

输出 发生在运维实体上的可疑事件

价值：快速确定影响范围，锁定怀疑范围

以技术方式对故障作快照

故障分析流程



类型	应用	事件	可疑程度
指标突变	淘宝/app1	RT突增	3
指标突变	支付宝/app2	QPS突降	2
变更	支付宝/app3	发布新版本	1

业务异常发现

如何获取拓扑？

查询纵向拓扑，获取可疑应用

查询横向拓扑，获取邻居可疑应用

什么是运维数据仓库？

查询运维数据仓库，获取可疑事件

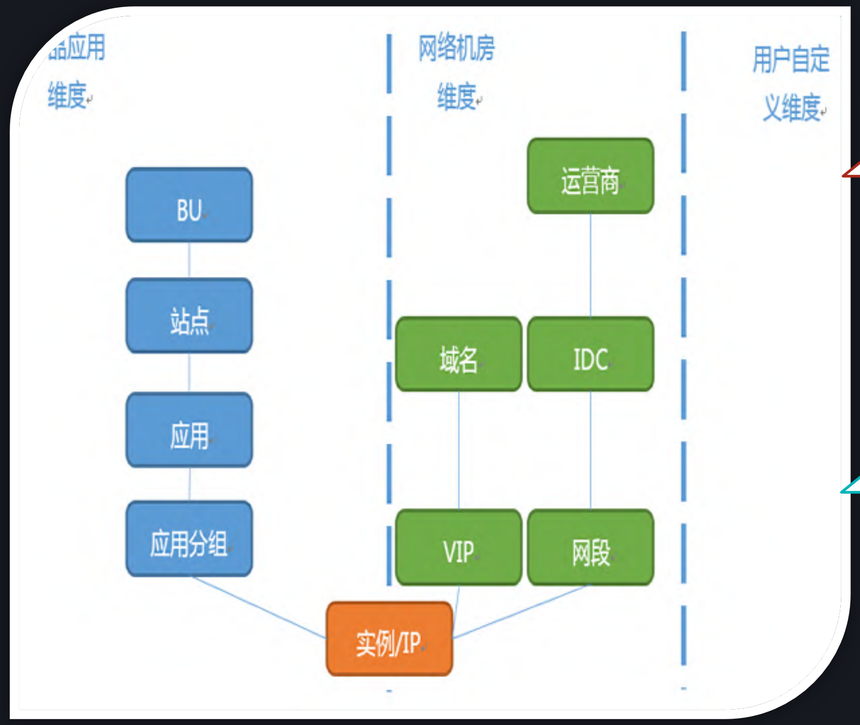
如何给出合理的排序？

根据故障定位算法，给出可疑程度排序

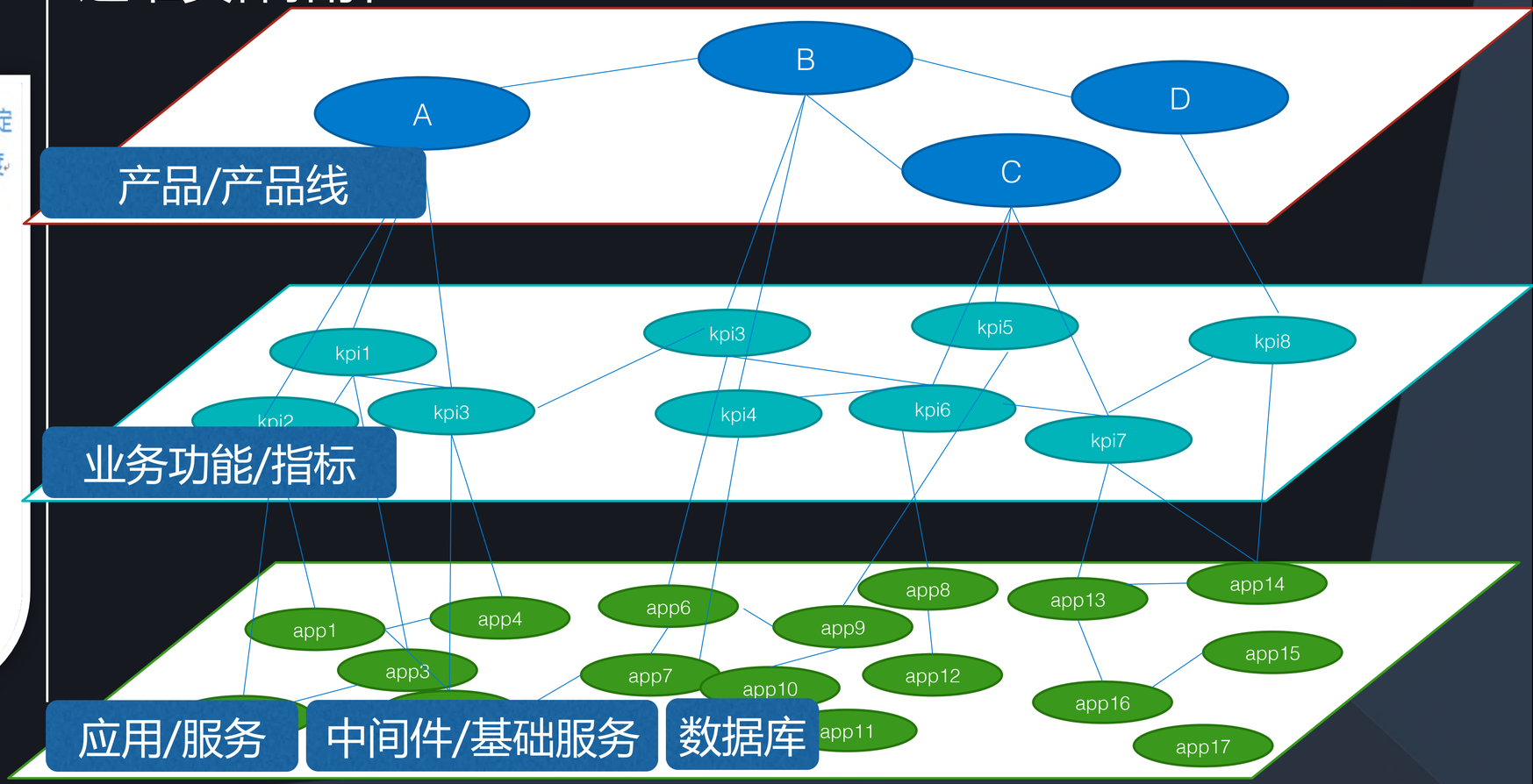
如何发现指标突变？

运维实体及其拓扑

运维实体维度和层次

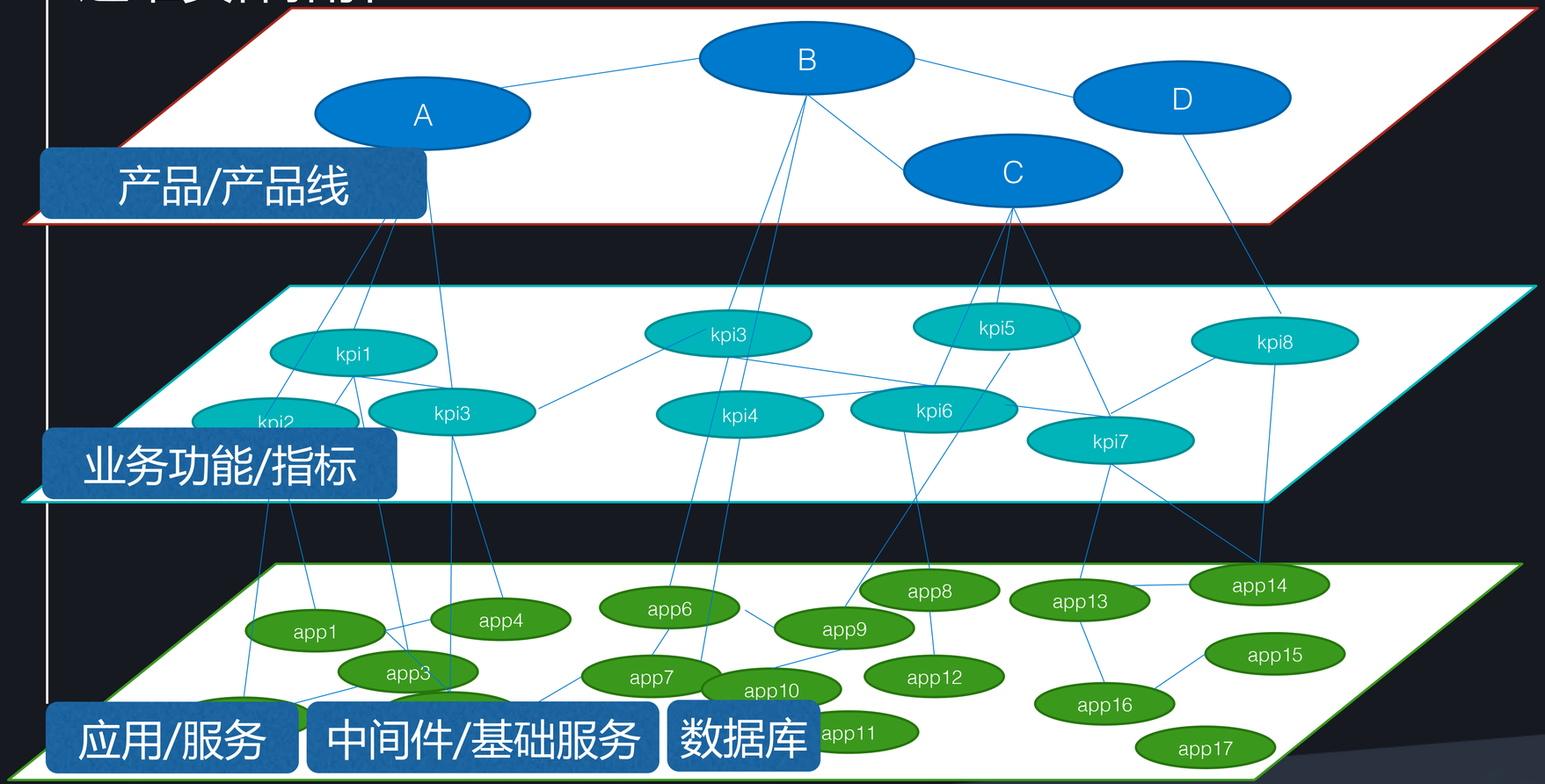


运维实体拓扑



如何获取拓扑

运维实体拓扑



横向拓扑

- 应用服务
  - TRACE系统
  - 手动配置
- 业务指标
  - 应用拓扑汇聚
  - 频繁项集发现
  - 手工配置

纵向拓扑

- 监控配置
- CMDB
- 手工配置

故障定位信息收集：运维数据仓库

业务功能

- 收集和故障相关的所有事件信息
- 提供按**运维实体**及其**拓扑**实时检索的能力
- 包括但不限于变更/上线，网络异常，系统/应用服务/业务指标异常，报警，日志异常等

技术架构

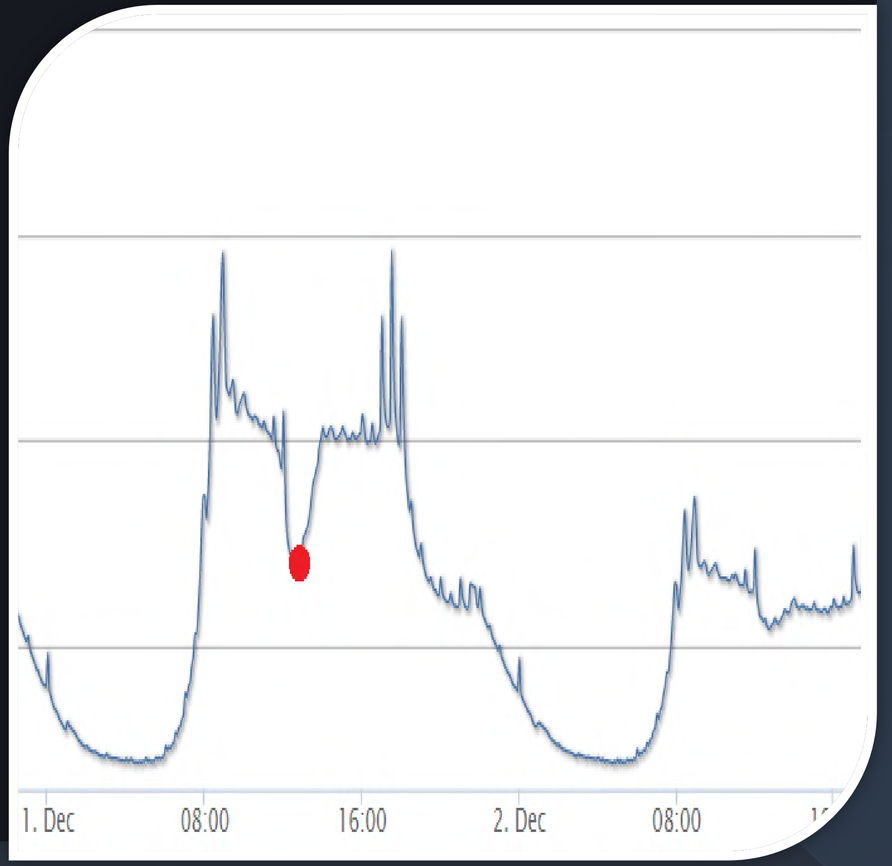
- 查询层
- 数据接入层
- 运维实体元数据管理/CMDB
- 拓扑管理
- ETL
- 在线事件存储
- 离线存储

技术化故障快照：

自动化检索和故障**相关**的所有**运维实体**及其上发生的**事件**

如何发现系统/应用指标突变？

异常检测	通过算法，自动发现系统/应用级指标中的跳变
	按需检测，由业务异常触发
为什么不使用监控报警？	覆盖率问题
	报警尺度不统一问题



如何确定可疑程度

对事件分类

- 现象类事件：指标突变、网络异常、日志异常
- 操作类事件：服务变更、网络变更

找到异常现象最显著的现象事件及对应的应用

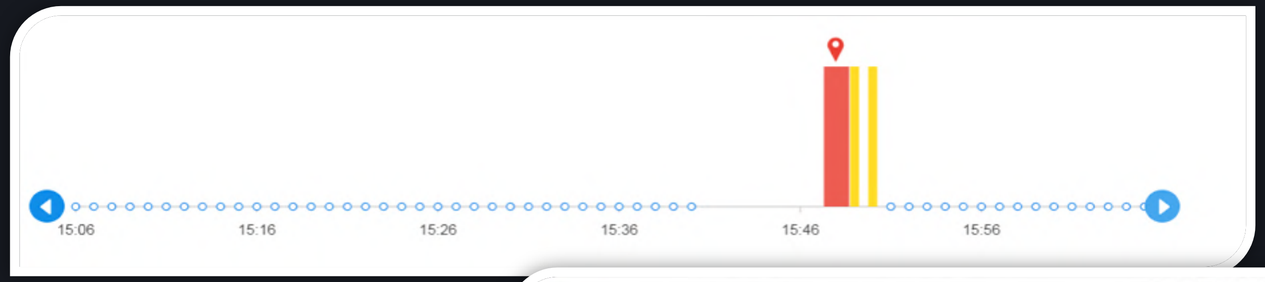
- 对突变指标的异常程度作排序
- 对其它现象加权排序

找到距离最显著现象最近的操作类操作

- 时间距离
- 拓扑距离

故障智能分析效果展示

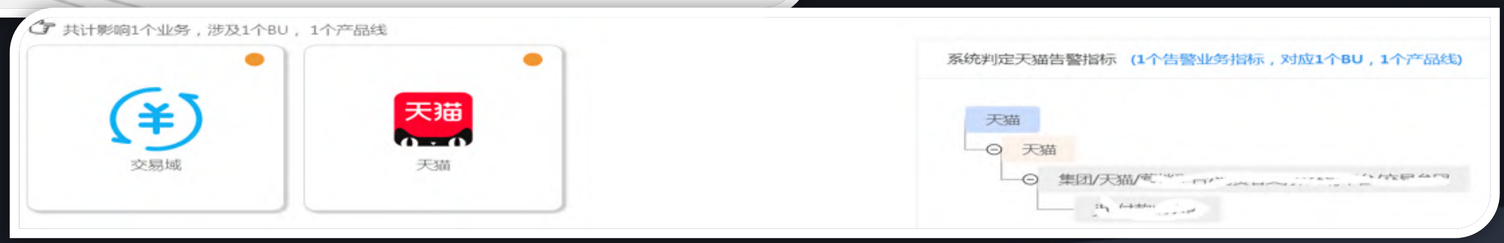
全局业务状态监控



1	2017-11-29 14:51	应用变更	中间件技术部/中间件/高可用架构	Date:2017-11-29 14:51:54,empId:76010,bcp规则配置变更:代码变更,id:3581
2	2017-11-29 14:50	应用变更	无线/手淘-微淘/手淘微淘	dataId: root group: quora_config units:[pre]
3	2017-11-29 14:50	应用变更	无线/手淘-微淘/手淘微淘	

相关可疑事件推荐

应用链路追踪



影响面实时展现

TABLE OF

CONTENTS 大纲

- 阿里巴巴全局故障治理流程和业务痛点
- 故障治理领域引入智能运维的效果
- 实战案例：业务异常发现
- 实战案例：故障智能定级
- 实战案例：故障智能分析的探索
- 智能运维项目落地的建议

智能运维项目落地的建议

从业务出发作问题拆解

- 流程问题
- 工程问题
- 算法问题
- 其它问题

重视数据积累

- 数据结构化
- 标注数据的质量
- 消除数据孤岛

关于算法选型

- 业务场景
- 数据的质和量

重视运营

- 量化、量化、量化
- 分类、分段分析
- 从一开始就关注业务结果

阿里巴巴官方首度分享

几十位工程师倾力总结技术实战经验

价值

现场签售

segmentfault

心作品

Alibaba Group 阿里巴巴集团 GOC Global Operations Center

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit 全球架构师峰会 2017