全球最大呼叫平台监控实践之路

ZABBIX 2019 Conference

王漫雪

技术经理，中移在线服务有限公司

1 背景-新挑战

2 反思-几个问题

3 沉淀-让监控多些可能

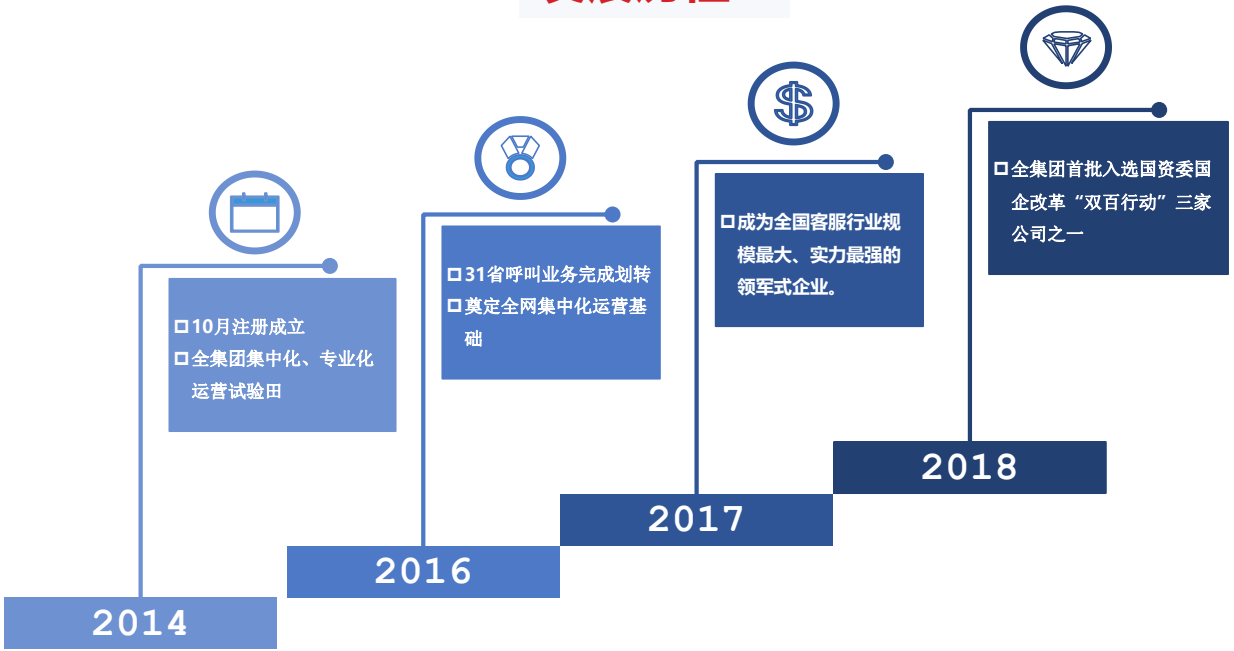
4 蜕变-AIOPS在监控报警方面的尝试

中移在线服务有限公司简介

公司服务宗旨

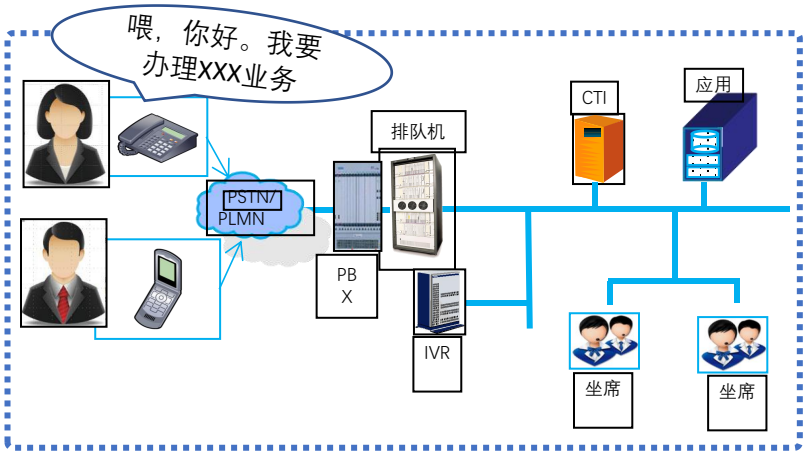
- 移动全网渠道运营集中支撑者
- 移动全网集中服务提供者
- 移动全网业务后台集中处理者

发展历程

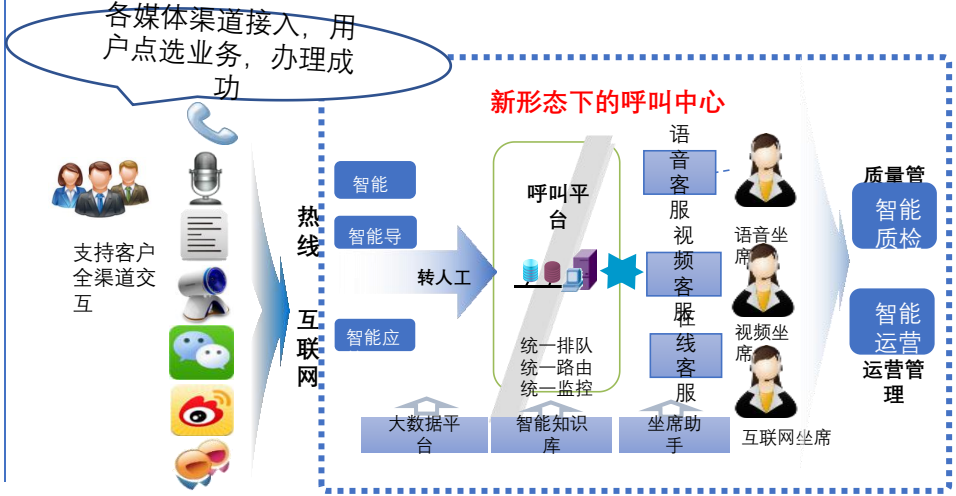


业务需求升级

你觉得服务场景只有这些



实际服务场景



业务需求升级

- 为满足多媒体渠道用户接入，公司进行了**服务、业务、运维**升级。
- **服务升级**：支持传统语音、文本、图片、视频、短语音、微信、微博等**多渠道、多数据**内容接入。
- **业务升级**：由软硬一体设备升级为**纯软件系统**，实现全媒体CTI、IVR、互联网接入网关、软交换、中继网关、媒体加速、用户终端**一体化运营**；将**人工智能（AI）、大数据技术**应用于IVR、机器人应答、质检、外呼等核心业务。
- **运维升级**：本部及31省分公司设备及业务系统，**集中化**配置、部署、上线、发布、迁移、维护、监控、告警、资源管控、故障处理。

积极应对挑战

面临的运维挑战

多

用户多，IT规模接近一线互联网企业

- 9亿 用户，超1亿微信粉丝，月服务超亿次，微博矩阵粉丝3038万（居行业首位），10086APP超五千万用户量
- 30000+服务器，15000+Pod
- 50000+Tomcat

难

业务变化快，运维环境复杂

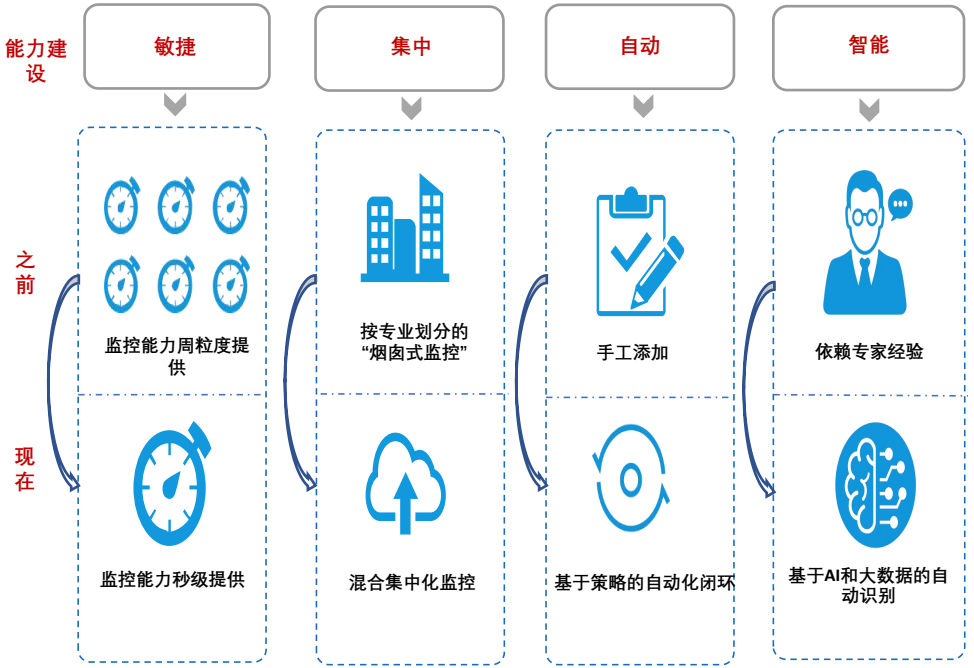
- 支撑全国营销活动，总部/分公司/省公司多级协同
- 日均上线 47 次，日处理 206 例工单
- 技术新：微服务/云计算/容器 ...

高

要求高，提供电信级服务

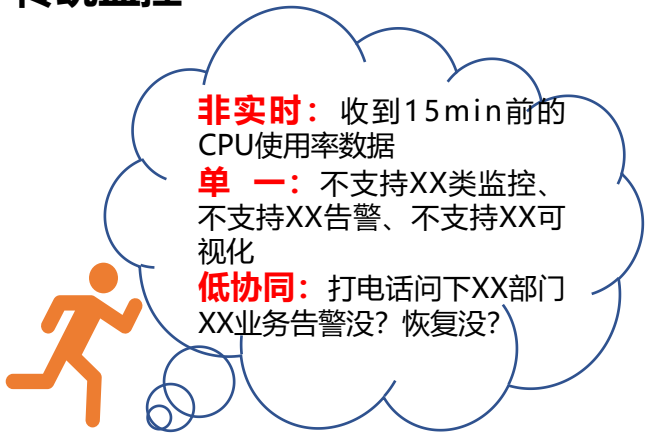
- 99.999% 的可靠性
- 15秒 接通要求
- 7*24 小时保障

积极应对



拥抱开源

传统监控



新时期监控需求

拥抱开源，站在巨人肩膀上

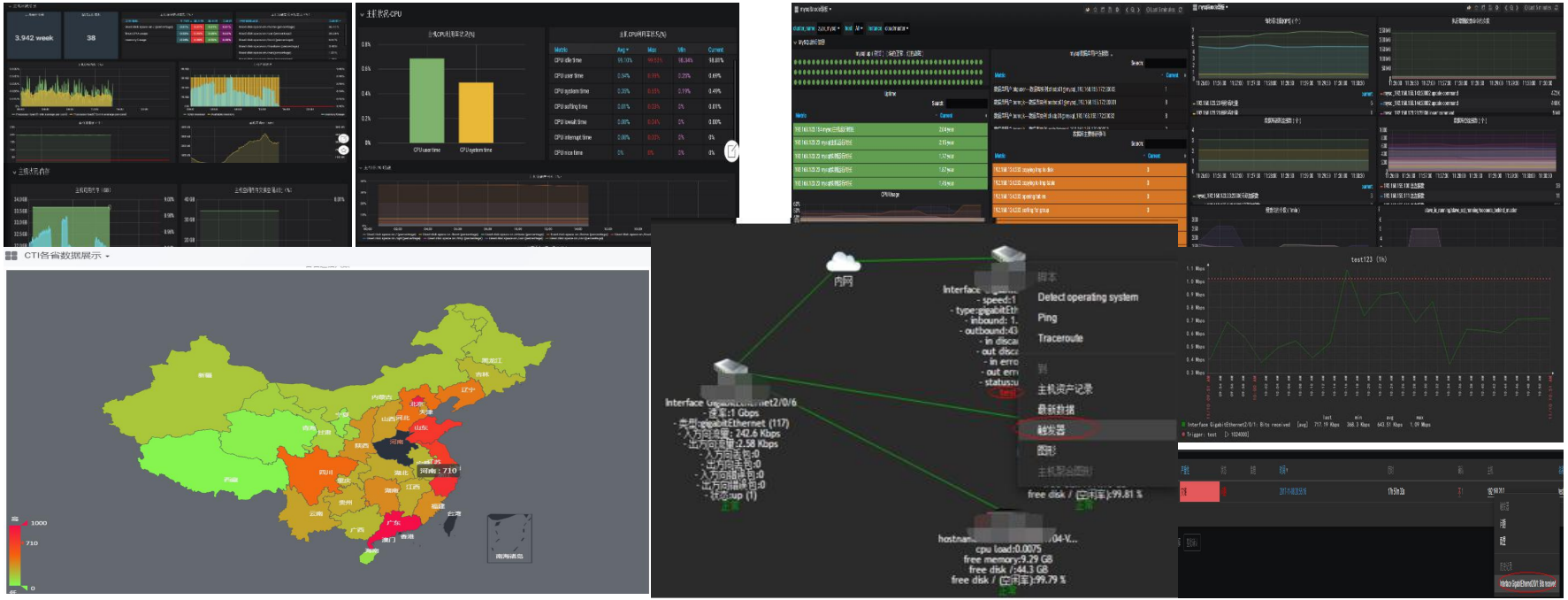
快速设施	利用Zabbix的成熟能力，1个月快速完成监控系统的能力建设
覆盖范围全	Zabbix自带的官方模板以及社区的各种模块，可以快速实现多操作系统、各主流中间件的监控覆盖
实时稳定	Zabbix非常成熟，可实现秒级数据采集，线上2年基本无故障发生
可视化看板	可与Grafana很好的结合，快速实现丰富可视化看板的制作
高效低成本	Zabbix本身资源消耗极低，主要是数据库需要物理硬件支撑，比Prometheus占用资源少

半年时间

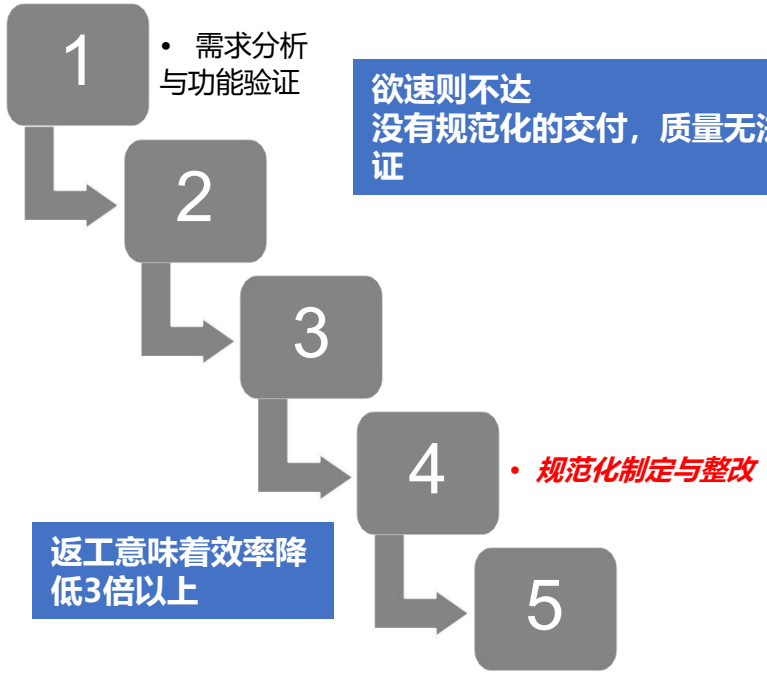
2 万	200 万	90 万	30 万
主机	监控项	触发器	报警
84	400+	545	1.3 K
Proxy	DashBoard	用户数	动作

一些展示

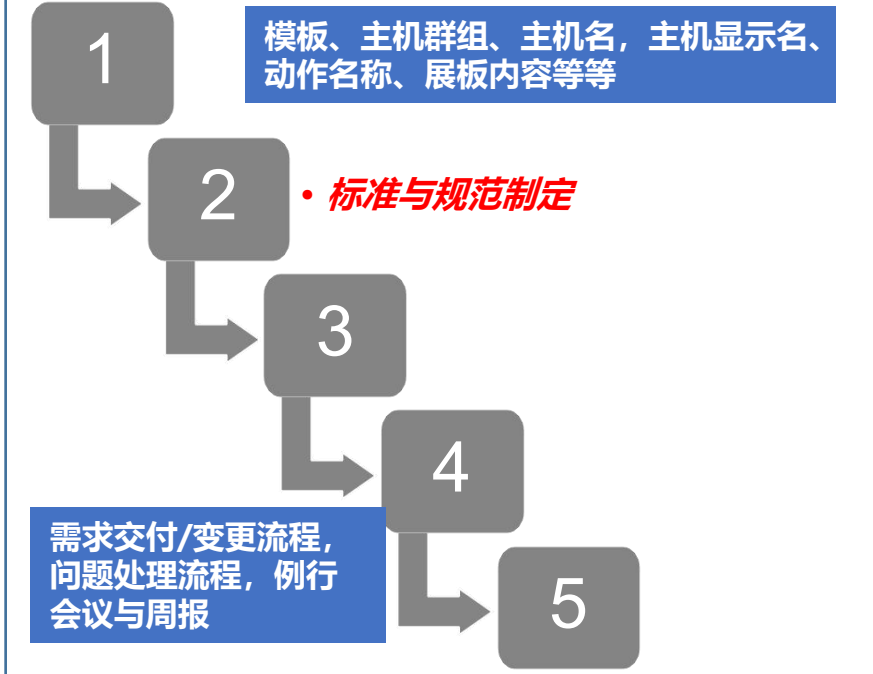
视图可灵活制定，可实现按照不同用户需求灵活定制，分钟级配置。可实现折线图、柱状图、饼图、区域图、拓扑图等多样图表。



中移在线监控的历程（摸着石头过河）



建议流程（标准先行，质量与效率并重）



中移在线

1 背景-新挑战

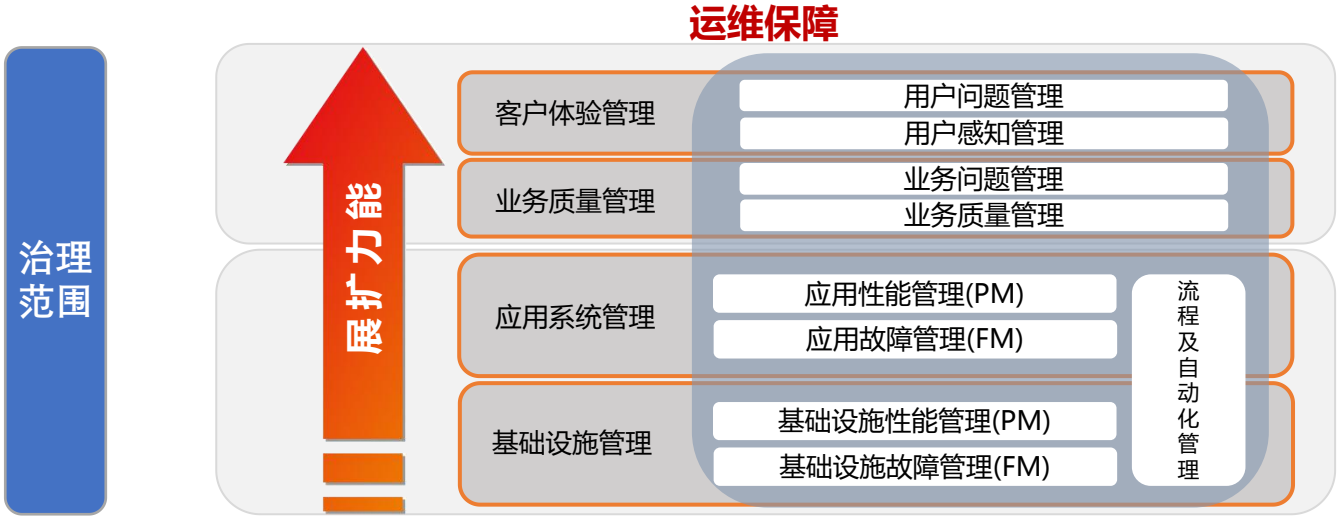
2 反思-几个问题

3 沉淀-让监控多些可能

4 蜕变-AIOPS在监控报警方面的尝试

问题一：200w监控指标，业务出了问题仍然不知道

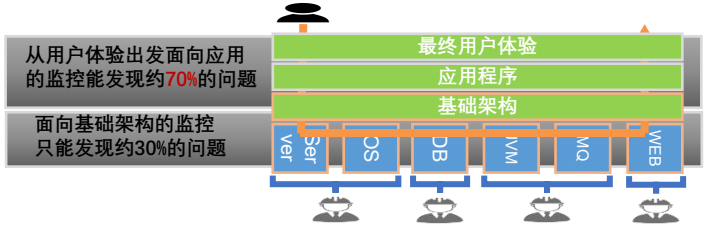
- 以业务质量和客户体验为核心，以可管控、可视化、可度量为目标。
- 全网集中建设、集中管控、边缘节点标准化接入。
- 软件监控+硬件监控一网打尽，运维数据统一、融合、流动，建立多层次度量体系。
- 以用户体验出发，建立端到端全链路监控，告警+投诉预警+客服联动形成完整闭环管理。



业务及应用质量可感知，是监控的核心

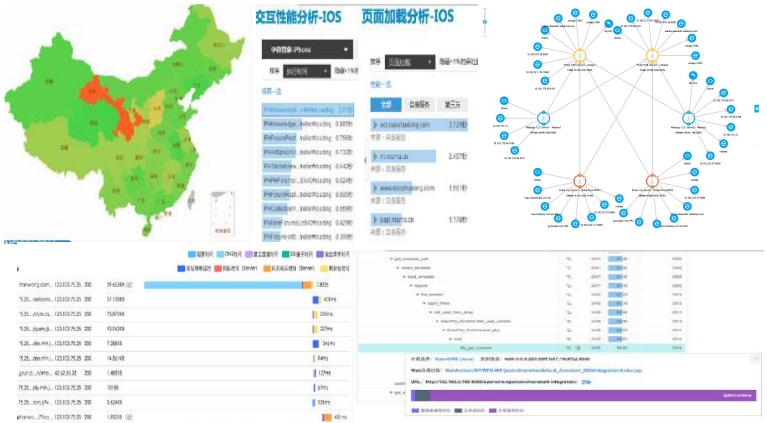
在强化基础设置监控的基础上，补充应用性能监控和业务质量监控能力，保障业务的稳定性和客户感知。

应用性能监控



应用性能监控将前台页面与后端服务以及用户网络环境真正串联，做到端到端全链路、代码级监控。

用户体验评分 → 前端交互体验 → 网络切片 → 应用调用拓扑 → 代码定位追踪

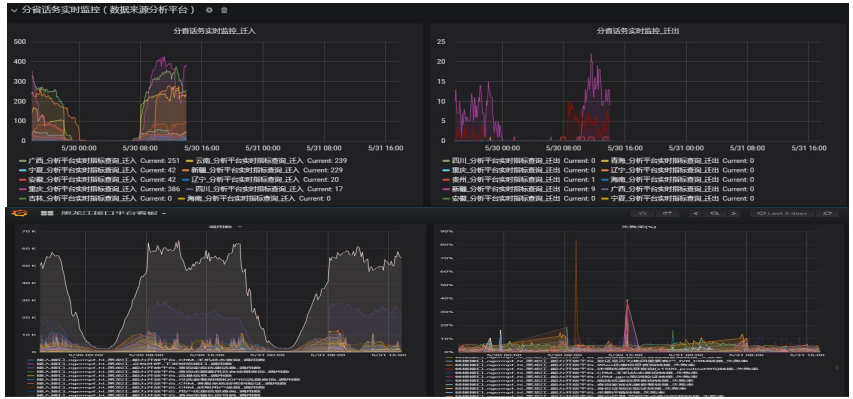


业务质量监控

梳理业务系统核心功能模块 → 梳理功能模块的核心监控指标 → 评审监控指标的提取方式及有效性 → 监控看板制作

参考Google SRE五项黄金指标

1：速率：请求速率，请每秒请求数量。
2：错误：错误率，即每秒错误数量。
3：延迟：响应时间，包括队列 / 等待时间，以毫秒为单位。
4：饱和度：即过载程度，指标与资源利用率相关，也可通过队列深度进行直接衡量。
5：利用率：资源或系统的繁忙程度，通常表示为 0% 至 100%。



问题二：海量的日志是否有利用价值?

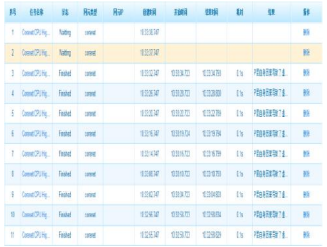
对于亚健康状态，异常日志比系统故障更早出现。由于海量日志存储在海量网元中，不同厂商日志标准不统一且可读性差，往往很难鉴别真正触发异常的日志。

挑战

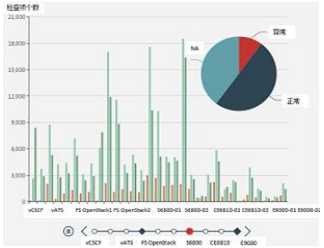
海量日志保存在海量网元中，缺乏统一视图

不同厂商设备的日志缺乏统一标准，可读性差

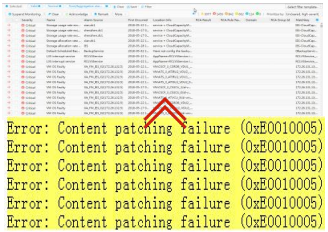
①跨厂商设备日志统一查询



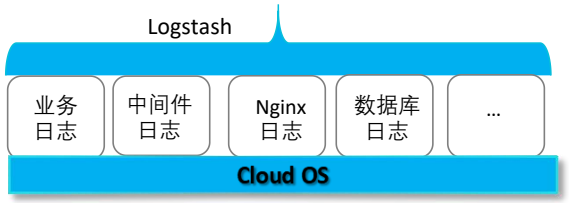
②异常日志统计

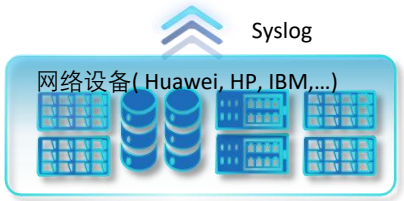


③异常日志分析与告警推送



统一日志分析



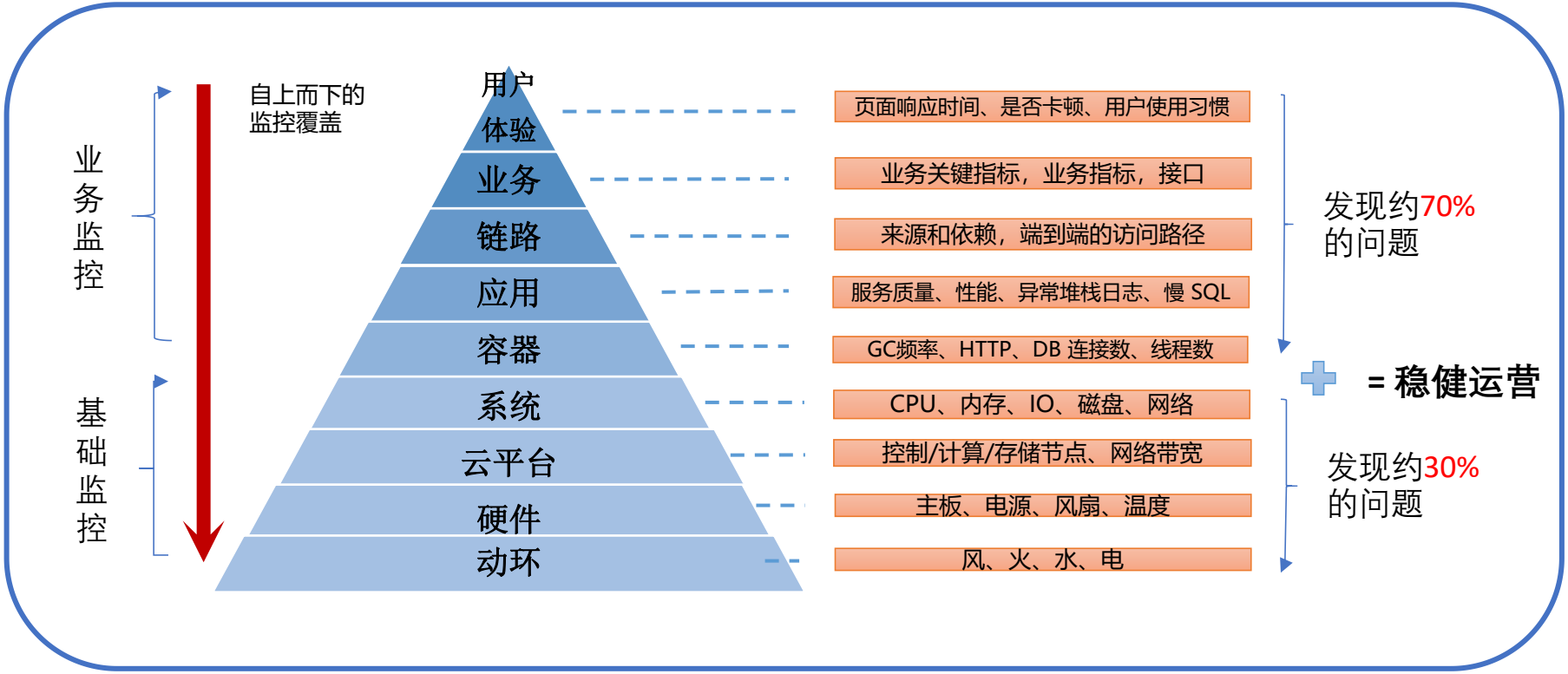


价值

日志统一采集，统一呈现，异厂商设备日志统一查询

针对异常日志进行统计，实时推送异常日志告警，提升亚健康网络问题定位效率

云化架构下的监控分层


业务监控
基础监控
自上而下的监控覆盖
用户体验
业务
链路
应用
容器
系统
云平台
硬件
动环
页面响应时间、是否卡顿、用户使用习惯
业务关键指标，业务指标，接口
来源和依赖，端到端的访问路径
服务质量、性能、异常堆栈日志、慢 SQL
GC频率、HTTP、DB 连接数、线程数
CPU、内存、IO、磁盘、网络
控制/计算/存储节点、网络带宽
主板、电源、风扇、温度
风、火、水、电
发现约70%的问题
= 稳健运营
发现约30%的问题

1 背景-新挑战

2 反思-几个问题

3 沉淀-让监控多些可能

4 蜕变-AIOPS在监控报警方面的尝试

自动化提速

引入自动化手段，封装标准模板，通过Zabbix和Prometheus的API调用，打通CMDB、监控、告警数据流，实现一键批量创建监控、告警的功能。

痛点

- 大型互联网公司基础资源多，业务广，线上变更频繁，监控配置任务量大
- 监控添加不是一蹴而就，需要反复调整，重复工作量大
- 开源工具使用门槛高，大多没有好用的web界面，需要培训才能灵活使用
- 中移在线公司业务/工作人员遍布全国各省，基础资源达到上万级别，业务变更频繁，统一管理难度系数高

应对方案

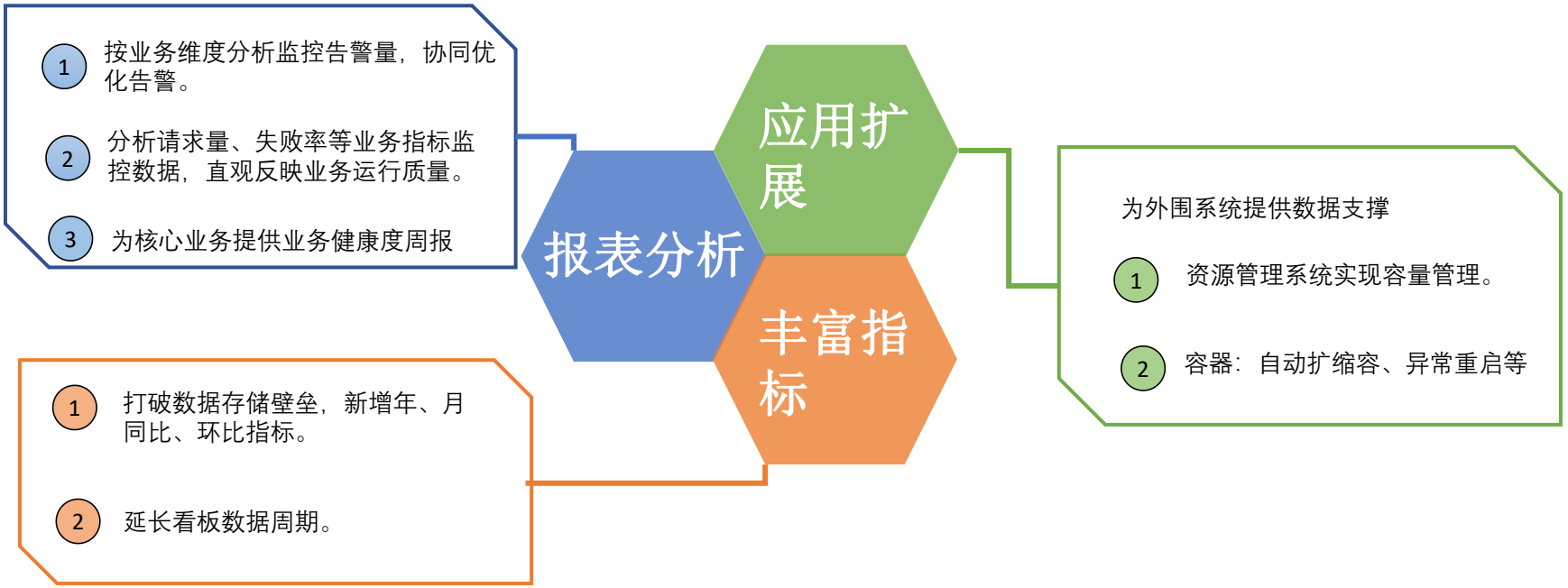
1. 监控能力流程化、模块化、标准化
2. 二次开发、自动化
3. 配置界面化 数据展示界面化

自动化提速

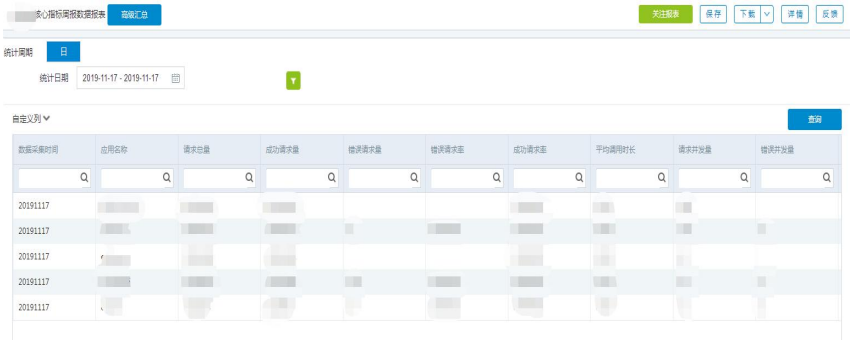
当前监控自助可实现各类监控自助式增、删、改、查，已覆盖公司**76.2%**的监控需求，目前只需4个自有人员即可维护整个监控系统。

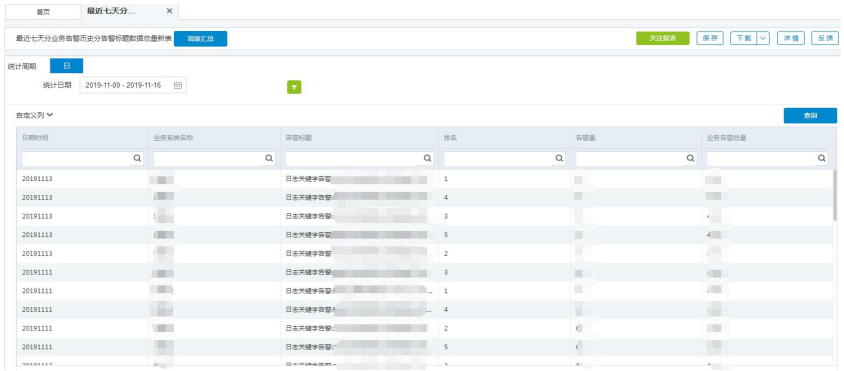
监控分类	监控对象
网络设备	华为、思科、锐捷、华三
主机	Suse、CentOS 、Redhat 、Windows、Ubuntu
数据库	Mysql、Oracle
中间件	Tomcat、Redis
进程	CPU、内存、存活、端口
日志	关键字
拨测	拨测状态
自定义指标	命令类、SQL类

通过ETL中的crossdata组件抽取zabbix、prometheus告警平台历史数据，最终装载到大数据分析平台中，进行多维度的数据分析。

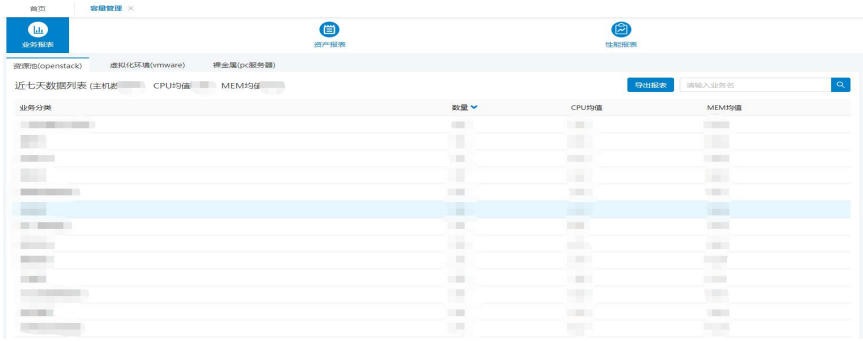


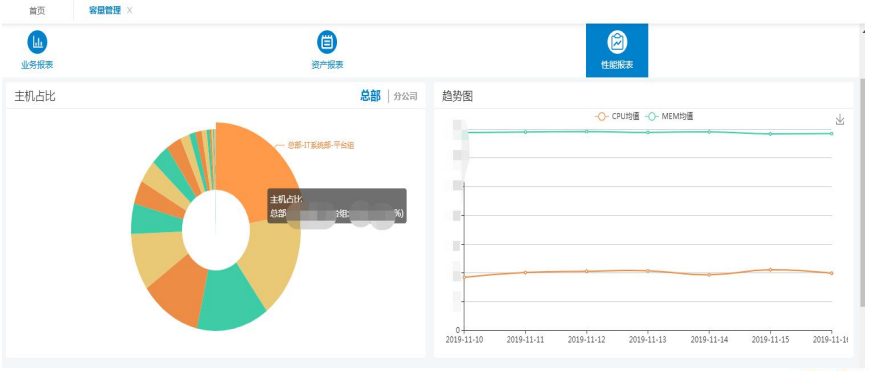
大数据分析报表





容量管理平台





现在的数字

3.3 万 主机	700 万 监控项	99 万 触发器	198 万 报警
220亿 日志	1300+ DashBoard	975 用户数	1.8 K 动作

1 背景-新挑战

2 反思-几个问题

3 沉淀-让监控多些可能

4 蜕变-AIOPS在监控报警方面的尝试

当前的主要矛盾：海量的告警和有限的专家

运维主管

工程师小明

监控要“多而全”，
一个问题都不能放过！

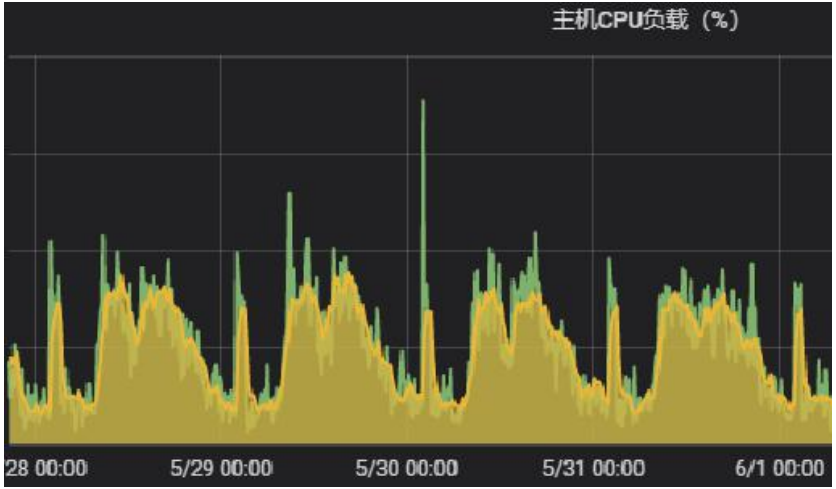
VS.

告警要“少而精”，
不要重复和误报

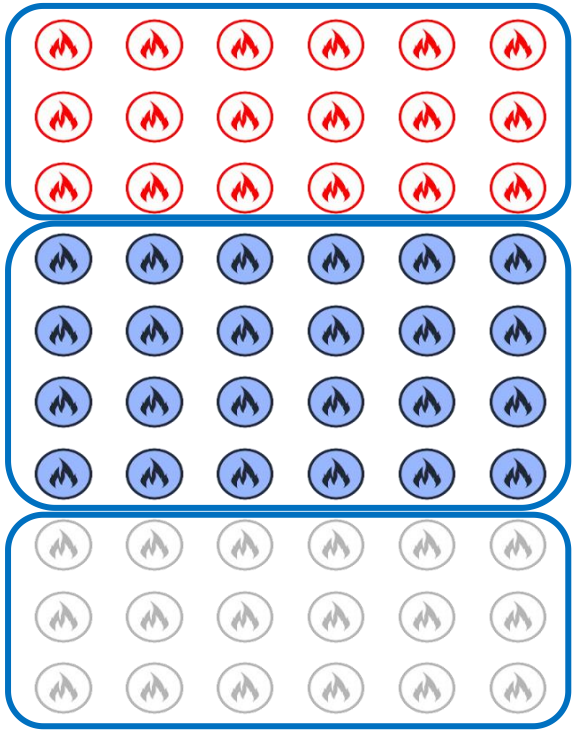
700万+ 监控指标，**99万+** 报警阈值，**198万+** 告警/天，**2000+** 短信/每人每天

一些问题

- 10086呼叫数量下降20%，是不是异常？
- CPU负载到了60%，是不是一定要告警？
- 海量的业务监控指标，业务部门无法确定准确告警阈值，运维部门如何设置阈值，才能同时避免误报（过于严格）和漏报（过于宽松）？



阈值正确设定是解决问题的关键手段之一



告警

- 正常告警
  - 缺少压缩&关联
- 误报
  - 阈值不合理：80%
  - 监控能力不足：10%
  - 人员配置失误：10%
- 漏报
  - 无法设定阈值：70%
  - 无监控：30%

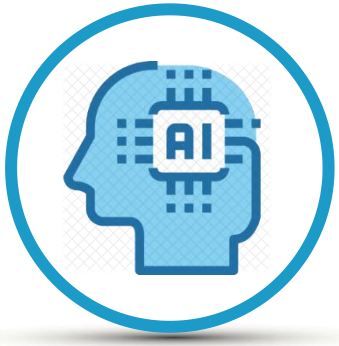
阈值设定从依靠专家经验向智能动态设定演进

专家依靠
经验设定

规则阈值

通过大数据
分析设定

固定阈值



通过智能分析
动态设定

智能动态阈值

基于结构化的时序数据，通过AI预测拟合曲线，进行异常检测

历史数据分析

历史数据读取和清洗

- 数据抽取ETL
- 断点修复
- 数据间隔调整
- 自相关性分析

毛刺检测

统计异常检测，用于过滤毛刺型异常

- Moving Average移动平均滤波 (ARIMA)
- Exponential Smoothing指数平滑滤波 (Holt-Winters)
- N*sigma统计检测

指标预测

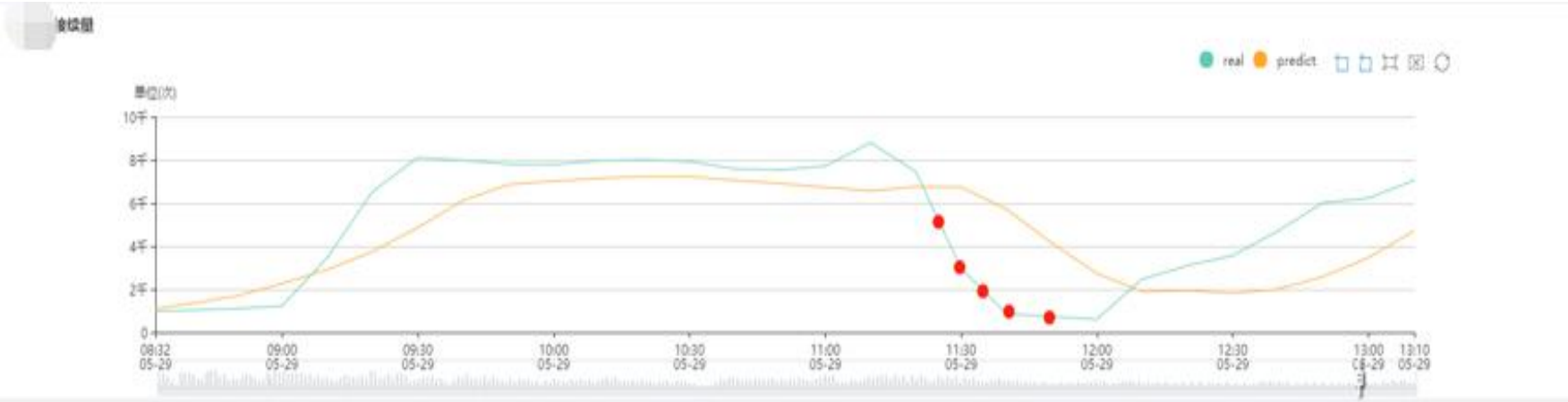
LSTM（长短期记忆）预测算法
孤立森林（Isolation Forest）
日同比（Day over Day method）
箱线图（Box-whisker plot）

异常判定

途径一：N-sigma方差

途径二：专家标记

预测结果与实际故障对比



智能化运维并不是我们想象的那样遥不可及

告警准确率提升到80%

告警覆盖率提升到95%

告警配置人力下降60%

数据

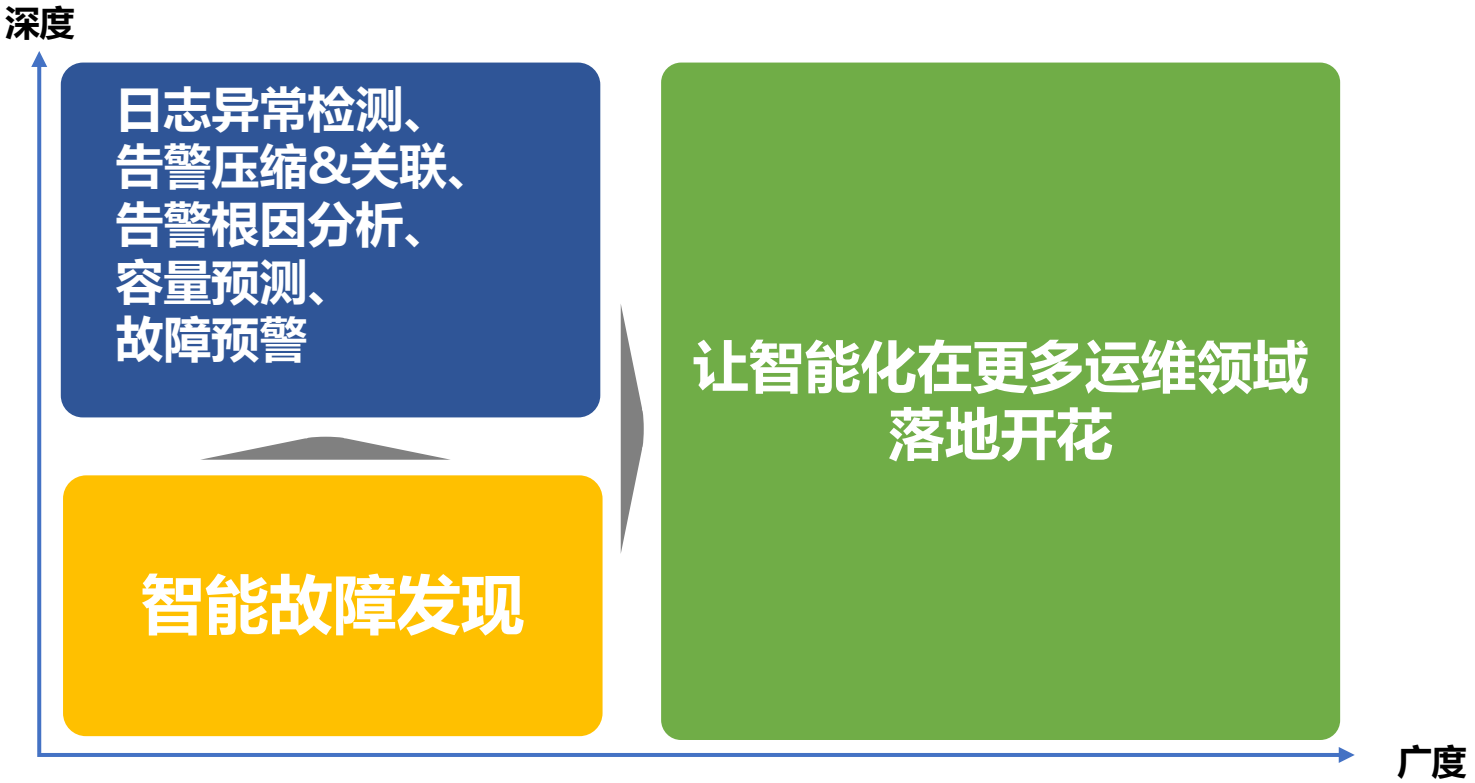
- 海量数据源（性能指标、日志、告警）
- 可以迭代预测、迭代标注……

算法

- TensorFlow等成熟算法库
- 针对不同场景，可选择不同算法，如LSTM用于趋势预测、ARIMA用于回归过滤异常

计算

- 轻量化
- 虚拟机部署，4C32G即可起步


深度
日志异常检测、
告警压缩&关联、
告警根因分析、
容量预测、
故障预警
智能故障发现
让智能化在更多运维领域
落地开花
广度

自动化提速

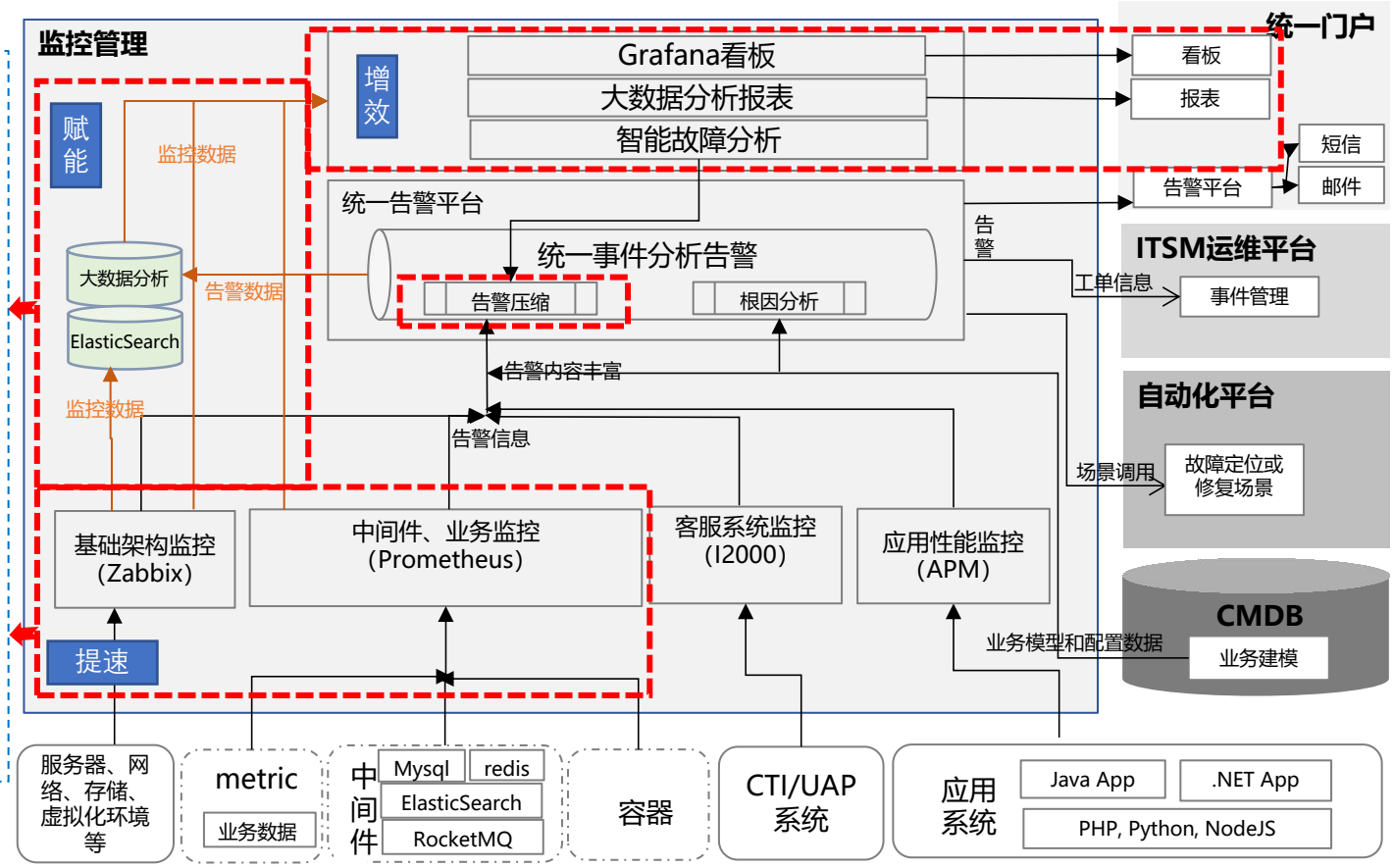
增强监控系统可用性，降低使用门槛，提速公司监控告警平台建设。

数据化赋能

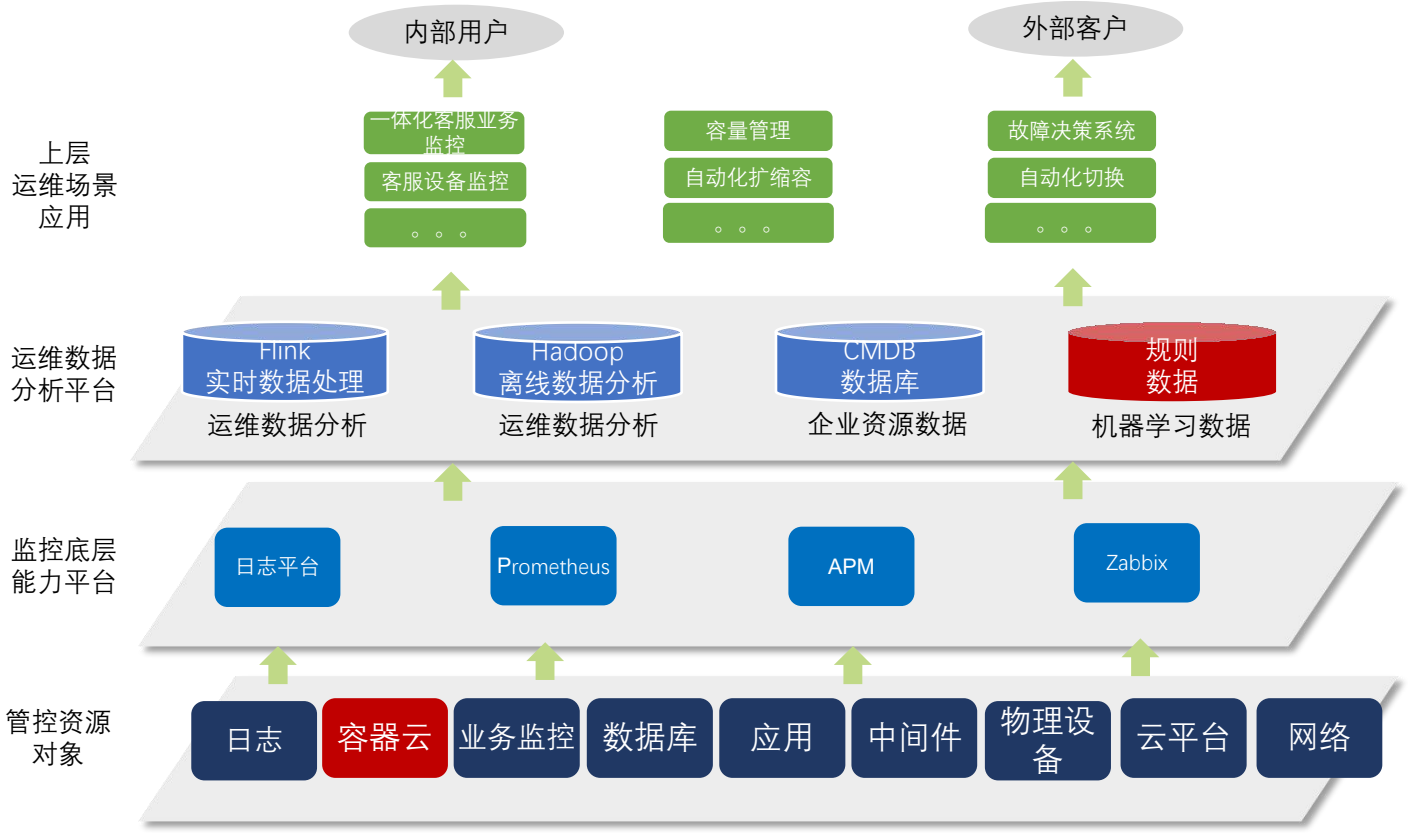
海量存储数据情况下，增加数据存储时长，达到丰富监控指标，扩展监控数据可用性目的。

智能化增效

建立丰富、多样、灵活的视图与报表，提供直观高效的巡检、定位工具。结合智能化手段，提升监控预警能力。



监控-总结


内部用户
外部客户
上层
运维场景
应用
一体化客服业务监控
客服设备监控
。 。 。
容量管理
自动化扩缩容
。 。 。
故障决策系统
自动化切换
。 。 。
运维数据
分析平台
Flink
实时数据处理
运维数据分析
Hadoop
离线数据分析
运维数据分析
CMDB
数据库
企业资源数据
规则
数据
机器学习数据
监控底层
能力平台
日志平台
Prometheus
APM
Zabbix
管控资源
对象
日志
容器云
业务监控
数据库
应用
中间件
物理设备
云平台
网络

加入组织

扫码入群

关注公众号

关注微博

联系我们


021-6978-6188


china@zabbix.com


www.zabbix.com/cn
www.grandage.cn


Zabbix开源社区

ZABBIX 2019
Conference
CHINA


Thank You!
ZABBIX 2019
Conference