云环境下的企业运维平台演进历程

阿里云 朱超健

阿里云 朱超健

十年互联网行业技术经验，专注于运维、安全、网络，具备丰富的运维平台产品建设经验；

早期就职于安全公司，经历了从传统安全模式到云环境安全体系的运维工具平台建设及落地；后就职阿里云，经历了云技术快速发展的关键时期，有从云平台底层到业务最上层的全链路专家经验，洞察运维平台关键点，长期专注智能运维领域，从事技术服务工作，聚焦金融、互联网、教育、泛娱乐等行业客户，基于客户业务打造托管式的云上智能运维解决方案，擅长用云最佳实践、产品管理、研发管理、业务重保、疑难问题攻坚等。

目录

- 多角度洞察运维痛点
- 企业演进过程中运维解决之道
- 阿里集团用云最佳实践
- 企业运维平台的未来

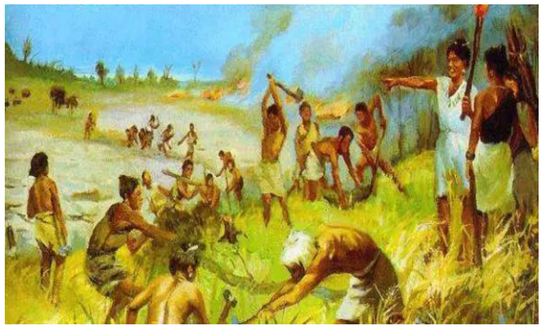
目录

- 多角度洞察运维痛点
- 企业演进过程中运维解决之道
- 阿里集团用云最佳实践
- 企业运维平台的未来

智能运维发展的必然性

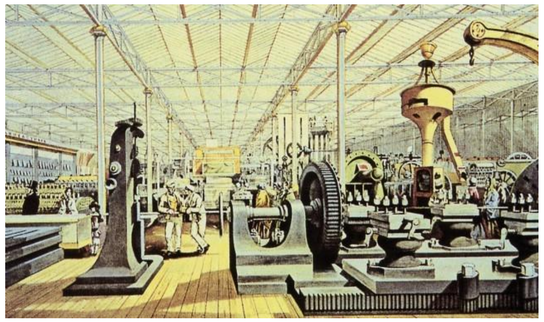
运维从人工到工具，从自动化到智能化也是互联网发展及企业业务发展的**必然趋势**。

农业时代



蒸汽时代

工业时代



智能时代

1.手工运维 → 2.脚本运维 → 3.自动运维 → 4.智能运维

手工运维阶段的特点——以人为主劳动，效率相对较低。因此，在这个阶段当企业IT系统发展到一定规模后，就会引发很多问题。正如生产力发展的农业时代。

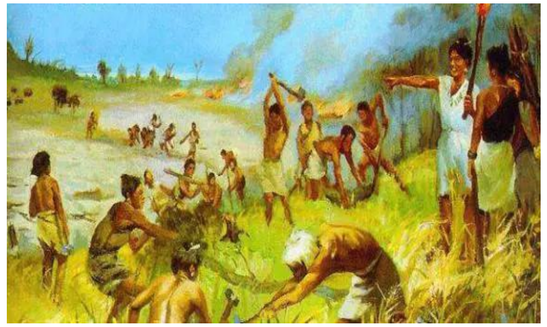
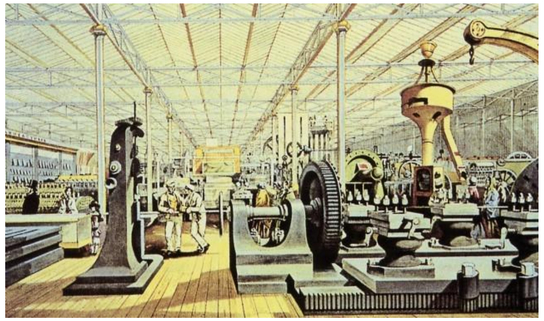
脚本运维，常常是运维人员通过实践沉淀了一小部分场景逻辑，使用shell来实现一小段简单的逻辑。只能说在手工运维的基础上做了简单升级，实则还有很大问题。正如生产力发展的蒸汽时代。

自动化运维工具和平台大幅度提升运维效率，让运维团队从机械、重复的劳动中解放出来。但随着运维工作的深入，自动化运维一些潜在缺点也逐渐暴露出来。正如生产力发展的工业时代。

在可以预见的未来，IT系统架构的复杂度越来越高，规模越来越大，同时伴随人力成本不断提高，渐渐地对于重型信息化企业来讲，运维不是简单依靠人力或传统的运维软件能解决问题了。正如生产力发展的智能时代。

智能运维发展的必然性

运维从人工到工具，从自动化到智能化也是互联网发展及企业业务发展的**必然趋势**。

农业时代	蒸汽时代	工业时代	智能时代
			
1.手工运维	2.脚本运维	3.自动运维	4.智能运维
运维资源不足	脚本适用范围小	问题判断依赖经验	机器学习
标准化程度低	无法自动运维	缺少数据量化支撑	根因分析
知识体系转移较慢	依靠人力运维	操作对事件影响不明	趋势预测
企业对IT系统依赖度高	运维稳定性差	知识复用性低	故障画像

目录

企业云上运维的核心问题

应用架构师　应用开发　云平台架构师　应用运维　CCB　专业运维

云上应用如何运维？

资源供给部署	逻辑态、部署态模型；流程编排；资源创建
监控与可观察性	监控指标，告警阈值，trace，事件定级
运维操作与自动化	扩缩容，巡检，备份与恢复，应急操作
可靠性与可恢复性	快恢，自愈，限流降级，容灾
合规与运维风控	风险识别、防护、检测、评估和处置

如何运维平台，如何保障业务稳定、高效运行，支撑企业稳定用云？

业务监控、智能基线、自愈能力

业务监控的意义

随着社会对互联网的依赖不断提升，互联网服务故障，影响越来越大

故障导致公司资金损失客户流失
还可能产生社会舆情及群体事件

摩菲定律告诉我们，如果一件事情有可能发生，那么必然将会发生，无法彻底避免

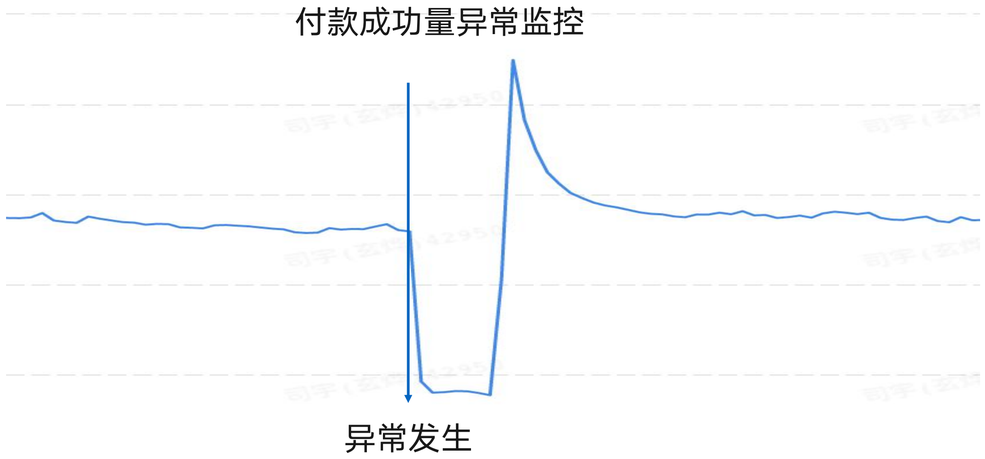
虽然故障无法彻底避免，但可以通过监控手段快速发现，缩短故障时长，降低影响

业务监控概述

历年故障分析

监控发现 **恢复时长远小于** 非监控发现

监控对于故障快速恢复非常重要



P1P2故障	故障发现时长	故障持续时长（MTTR）
监控发现	分钟级	分钟级*4
用户上报	>小时级	>小时*2

为何选择**业务监控**发现故障？

- 相比其他监控，对公司核心业务指标进行监控，更易发现业务异常

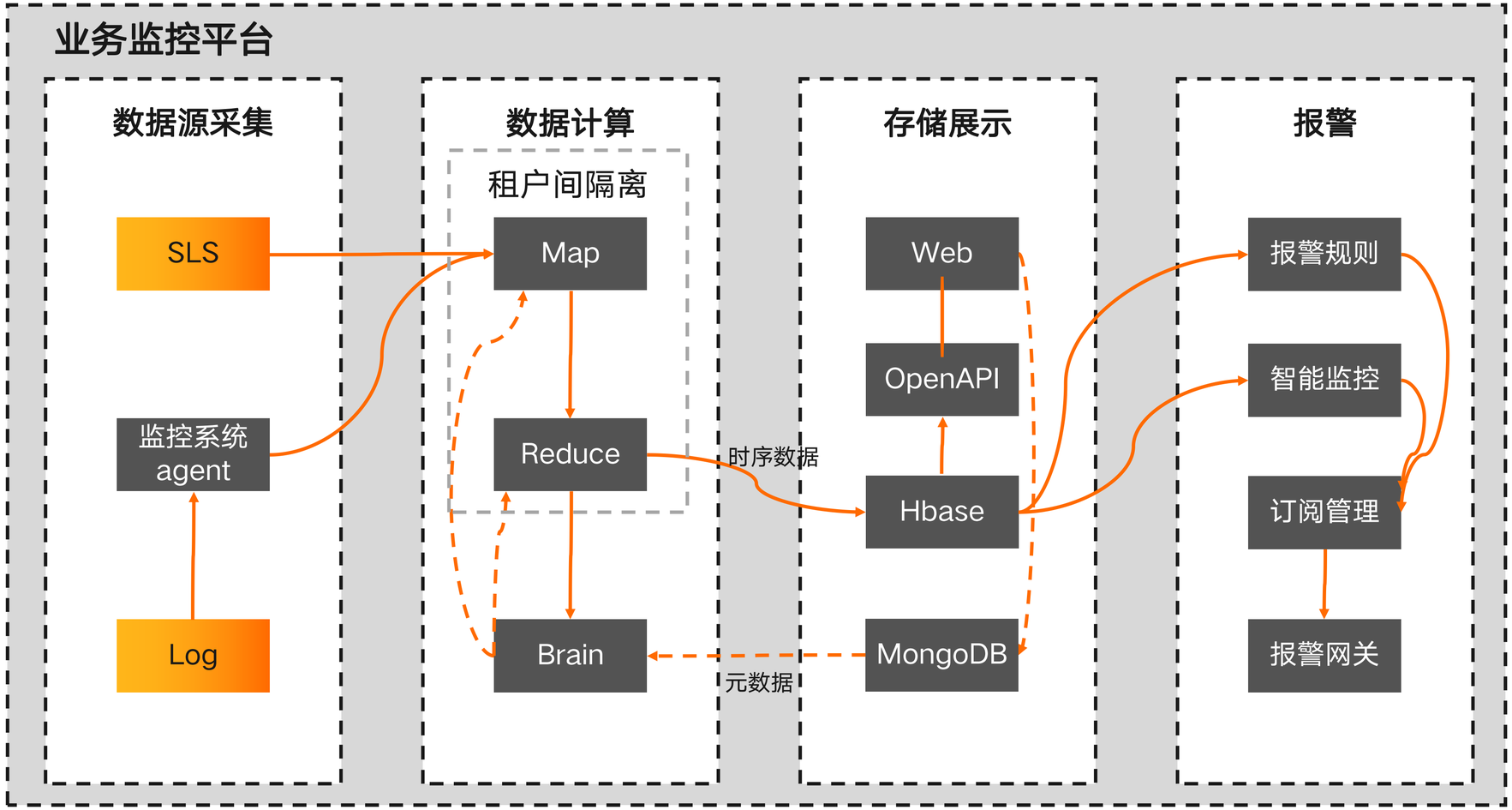
业务监控选择

~~SDK~~ ~~Export~~ ~~拨测~~ … 日志

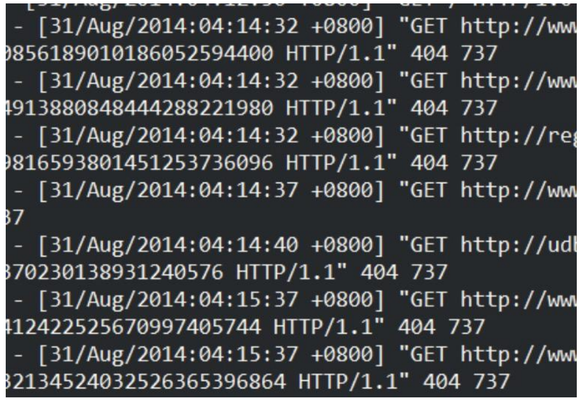
- 代码无侵入
- 业务改造少
- 信息详细

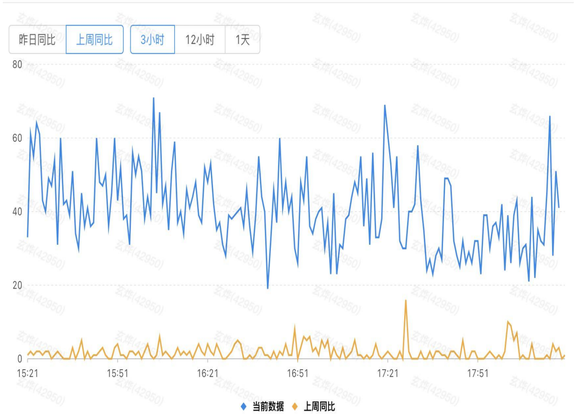
业务监控技术方案

- 调度10万+核计算资源，提供分钟级百T日志处理能力，存储亿级监控项



原始日志





时序监控

业务监控-指标自定义采集配置

可视化配置，将非标日志转换为统一时序监控数据

	名称	路径	环境	详细筛选						统计
说明	简单直观的反映监控指标代表含义	• 日志文件完整路径 • 增量采集	机房/IP/预发/线上等	压测标	业务逻辑	业务结果	响应时间	错误码	列-多维度	求和、平均等
				• 关键字段前后有明确标识分割（如 | ） • 以上筛选能力一般是基于具体业务需要场景组合使用						

业务监控-自定义报警规则配置

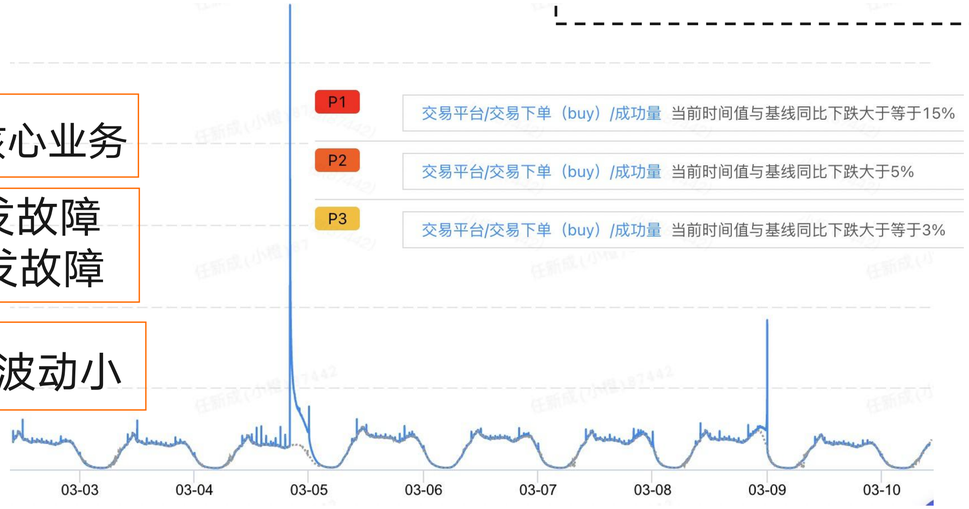
	淘宝 交易创建 – 量大稳定	菜鸟 仓储操作 – 波动稳定	饿了么 退款申请 – 量少
业务特征	• 量大，周期趋势稳定 • 故障等级：异常下跌5%触发故障 • 场景复杂，成功明确，失败可能非系统异常	• 量中等有抖动，周期趋势稳定 • 故障等级：异常下跌15%持续3分钟触发故障 • 场景相对简单，成功失败原因均明确	• 量少，无周期趋势 • 故障等级：持续5分钟成功率低于80% • 场景相对简单，成功失败原因均明确
报警规则	• 成功量8-22点 环比下跌超过2% • 或 成功量 22-8点 2分钟求和环比下跌超过3% • 不宜配置失败量报警	• 成功量 3分钟求和环比下跌10% 且 成功量3分钟求和昨天同比下跌10% 且 成功量3分钟求和上周同比下跌10% • 或 成功率 持续2分钟小于95%	• 成功率 持续3分钟低于90% 且 失败量 持续3分钟大于等于5 • 或 成功量 持续3分钟为0（兜底入口异常无失败量） • 不宜配置成功量波动报警

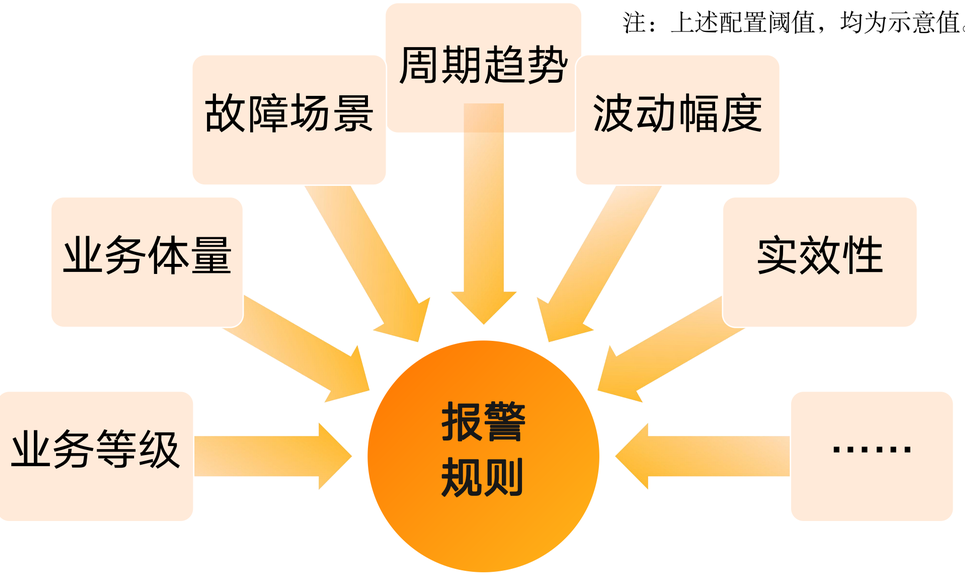
注：上述配置阈值，均为示意值。

业务等级高，全站交易核心业务

故障场景下跌3%就触发故障
实效性高，1分钟就触发故障

业务体量大，趋势稳定波动小





智能监控-智能基线

智能基线——基于机器学习算法的业务监控无阈值异常检测

实现方案

基线拟合	STL
异常判定	N-sigma

面临挑战

不同业态曲线的特征有较大差异

- 数量级
- 局部波动程度
- 周期

不同业态的异常判定标准有较大差异

- 线上业务（游戏）.vs. 线下业务(新零售)

优势

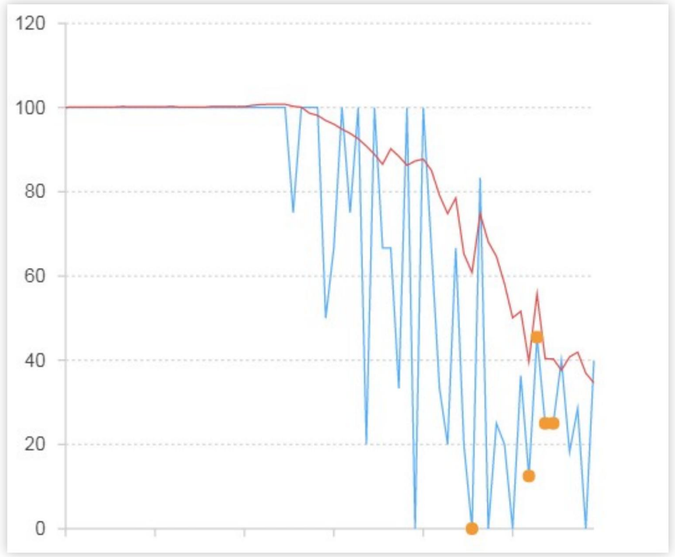
无阈值
高召回
低误报

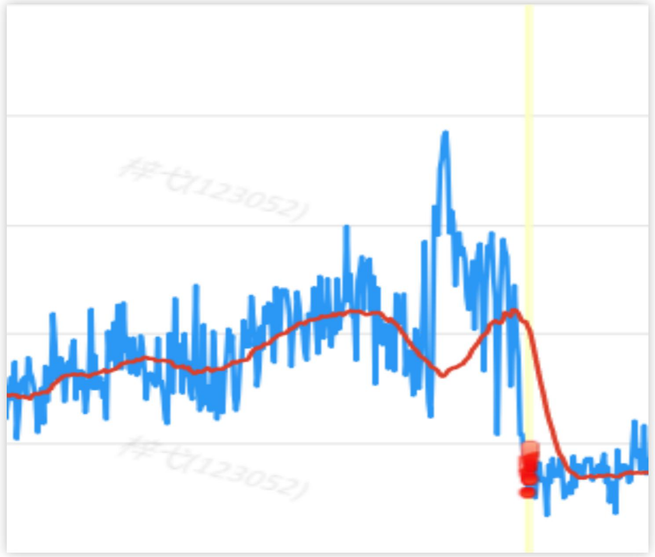
对外部干扰抵御较差
非周期曲线支持不足

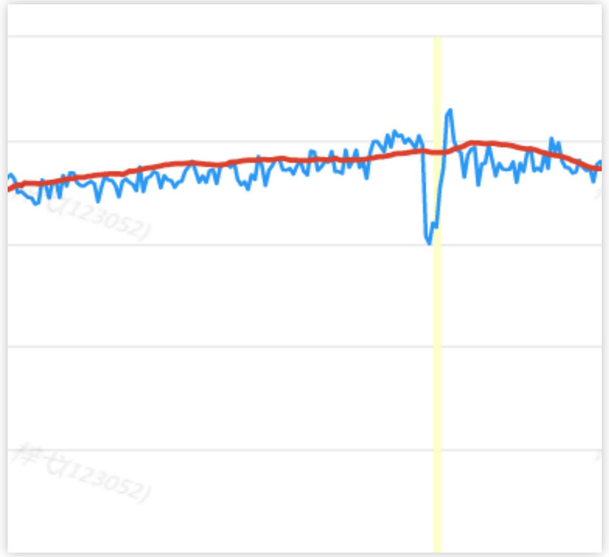
不足

智能监控-实际效果

针对周期性特征明显的业务监控指标，具有较高召回率及准确率







准确率>70%

召回率>90%

监控报警核心指标

通过核心指标，衡量公司业务监控报警质量，降低故障影响

准确率

取决于监控报警配置质量
准确率低误报多
导致报警成本高

报警准确率>=50%

01

召回率

取决监控覆盖及报警质量
决定异常是否能被监控发现

重大故障 >= 90%
一般故障 >= 70%

02

报警量

取决于合理订阅及准确率
报警量大
员工无法有效应急

人均日报警<=30条

03

目录

- 多角度洞察运维痛点
- 企业演进过程中运维解决之道
- 阿里集团用云最佳实践
- 企业运维平台的未来

阿里巴巴上云过程

阿里集团已经实现业务跑在云上，从回顾这几年上云过程，基本可以划分三个阶段。

- **上云历程**

集团上云三个阶段：**弹性上云 ➤ 核心系统上云 ➤ 全面上云**

在每个阶段集团上云解决的问题和核心关注点都是有差异的。同时在每个阶段达成具有里程碑意义的上云案例。

2017~2019 弹性上云

连续3年完美支撑双十一购物节上阿里云

神龙计算架构通过双十一验证，并在公共云对外商业化

2019~2020 核心系统上云

电商核心系统全部上云

计算平台搜索广告等业务中台上云

阿里云运营支撑上云

充分利用公共云的弹性，云上降成本

2020~ 全面上云

经济体一环/二环BU增量业务全战上云

考拉 饿了么 高德 优酷等100%上云

统一资源池，电商搜索计算平台混部

中间件云产品化支撑集团上云

极致弹性，SP模式引入集团上云

用云优化，云上降本提效

云上冬奥业务连续性保障方案

业务连续性保障方案 - 风险治理、容灾演练、压力测试、安全加固、预警风控、应急预案

风险治理

- 基于飞天技术服务平台（Apsara ServiceStack）CloudDoc/Advisor模块能力进行云平台风险巡检并前置治理风险。
- 重保期云平台针对性封网管控及变全网更评审把控。
- 云平台集群水位评估与管控。
- 批量资源预留和资源腾挪。

识别云基础设施潜在风险

容灾演练

- 云平台基础设施容灾能力验证，如负载均衡SLB 多可用冗余验证，RDS数据库HA切换验证等。
- 业务整体架构容灾演练验证，如跨域专线。
- 业务系统容灾演练：演练业务损失某单元模块功能的系统容灾切能力。

保证关键系统高可用

压力测试

- 利用单元压测摸排各模块性能瓶颈，并完成容量评估。
- 利用全链路压测方式验证系统整体并发能力是否符合业务需求。
- 对系统全链路性能瓶颈点做性能调优。北京冬奥累计压测奥运相关项目数十个子模块，数百个接口，性能优化2-6倍。

保证关键系统并发性能

安全加固

- 数据中心建设期间，2019年开启安全架构和策略设计。
- 大型国家级安全攻防演练。
- 主管单位、冬奥组委和各厂商安全情报协同处置。
- 冬奥重保期间蜜罐捕获请求数千次，恶意请求拦截超千万次，云安全中心告警及处理超千次，并封禁大量恶意IP。

构建纵深防护体系

预警风控

- 钉群机器人主要产品核心告警项目35项，重保期核心告警主动处理41次，避免风险扩大。
- 利用资源Grafana监控大屏按照top异常资源观察和汇总异常实例资源信息，做到全局实施观测实时处置。

识别赛事风险并处置

应急预案

- 按问题场景梳理准备应急预案73项，覆盖云上弹性、网络、安全、数据库、容器、存储、大数据和中间件等8个产品垂直线方向。
- 覆盖过载、丢包、业务IP错误拦截、黑洞清洗、超限和管控异常等不同问题的应急处理。
- 主要产品钉群机器人35类核心告警处理预案。

赛事问题快速恢复

账号与权限治理

基于业务和组织进行云上资源的身份管理和授权规范

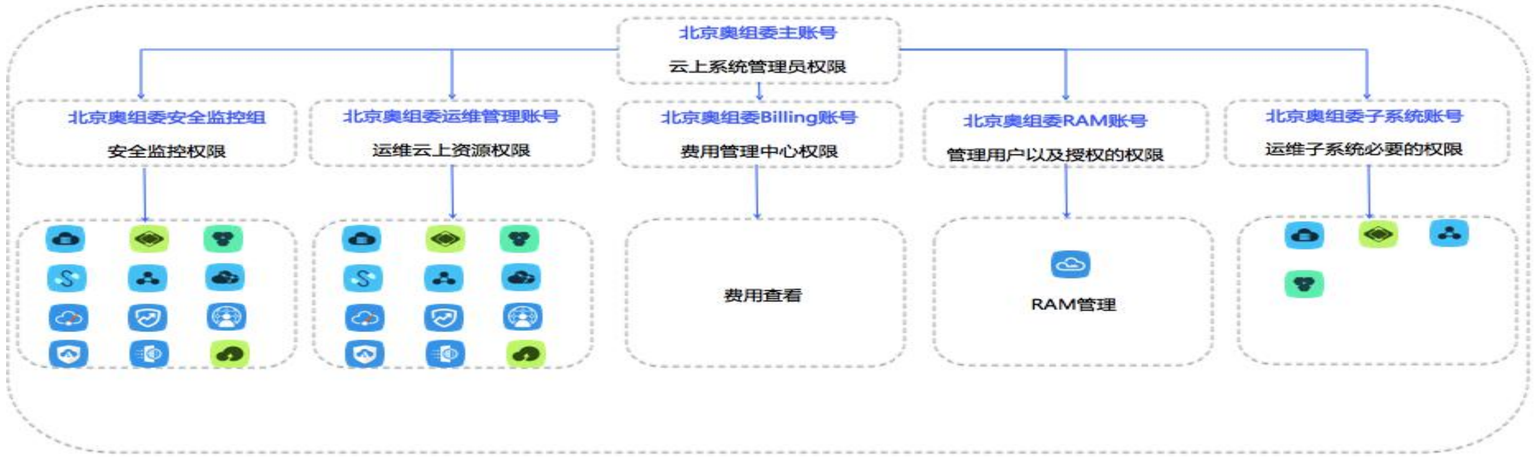
治理原则

- 根据组织架构对云资源进行分组，并以云上用户组为最小粒度授予相应的权限
- 权限越界访问用户数审计
- 授权失败审计
- 未在指定时间登录的用户数审计
- 闲置策略数审计
- 未配置强制多的因子认证的用户数审计

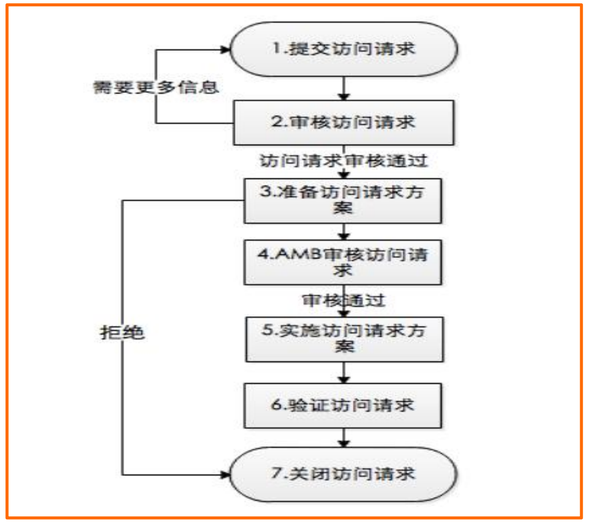
治理实践

- 根据组织架构对云资源进行分组，并以云上用户组为最小粒度授予相应的权限
- 云上用户组在信息系统运行不同阶段权限的治理
- 制定访问管理流程
- 配置审计进行持续合规审计
- 云安全中心进行持续合规审计

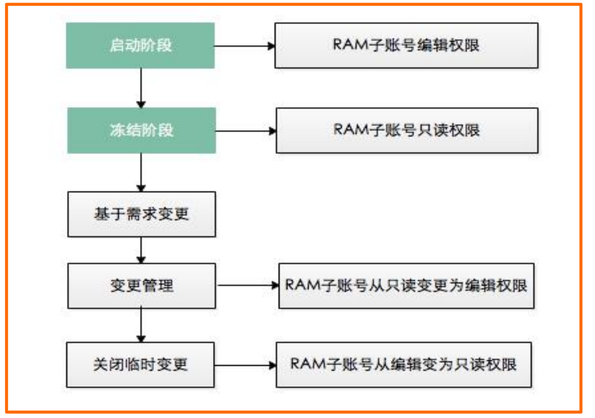
Beijing2022 总体账号设计



用户权限访问控制流程



信息系统运行不同阶段权限治理



资产与数据安全治理

基于丰富的阿里云安全标准化产品巡检发现问题，进行资产与数据的安全治理

治理原则

- 所有已部署的资产必须按照重要程度和数据敏感性进行分类
- 在可以批准并实现足够的安全和治理要求之前，无法将任何使用受保护数据级别的资产部署到云
- 任何包含受保护数据的段中提升权限都应属于异常
- 定期检查可能影响云部署的趋势和攻击，以更新云中使用的安全管理工具
- 日志按需持久化便于进行安全溯源

治理实践

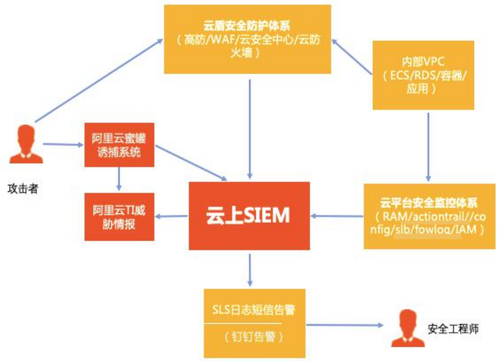
- 依据安全法规、冬奥业务类型，对资产和数据进行分类。
- 依托SSL保证数据传输过程的安全性，依托KMS、加密服务对敏感数据进行安全存储
- 基于操作审计、配置审计、日志监控一体化系统、数据库审计、堡垒机审计等对数据提取操作进行合规审计
- 定期检查可能影响云部署的趋势和攻击，不断优化安全工具

Beijing2022 云上安全产品

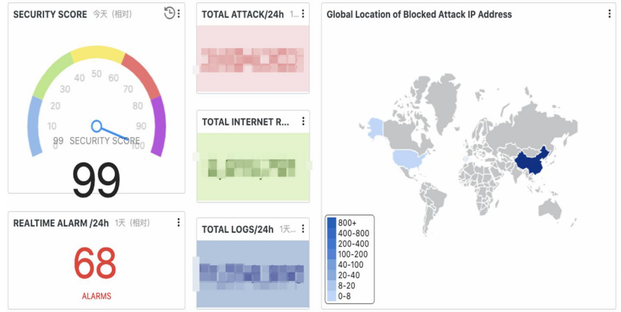
安全产品
WAF
Anti-DDoS（BGP）
Anti-DDoS(国际)
加密服务
堡垒机
云安全中心
云防火墙
日志服务
KMS
人机验证
数据库审计

SLS Audit Center

云上威胁一体化监测方案



安全监控大屏



全过程安全防护保障

基于强大的安全服务团队保障赛前赛中赛后的全过程零安全事件

历经4年构建了严密的纵深防护体系（ 应用、网络、主机、蜜罐、综合分析、预警 、架构设计），赛前主动解决产品稳定性风险；推动WAF接入率到达 97% 以上，实现高防和云安全中心100%覆盖率，默认重保模式，修复安全漏洞，大型国家级攻防验证10+，梳理应急预案30+，确保赛前风险最小化。

蜜罐捕获请求数千次，恶意请求拦截超千万次，云安全中心告警及处理超千次，封禁大量恶意IP，联合网信办、冬奥组委会处置、同步威胁情报。

全生命周期默认安全建设	**整体默认安全架构设计，默认deny策略，尽早发现安全防护的缺陷，通过服务弥补产品不足**
全面风险评估和安全验证	**多轮125项的全面风评，赛前完成安全加固，经过N次内外部的攻防演练验证安全防护水位**
安全责任有效区分落地	**组委、三方厂商之间明晰责任模型，确认底线并且形成文档，明确自身防守区域和权限范围**
情报协同与应急预案	**情报协同与应急预案：协同IOC，网信办，三方厂商的安全信息，以便及时响应并对外发声**

云上冬奥保障工具平台支撑

数十个云上系统的数百个关键指标的全景观测、告警，问题诊断与快速恢复

云基础设施全景大屏

全景作战大屏

业务信息系统全景大屏

飞天技术服务平台 - CloudDoc/Advisor

交互	控制台	OPEN API	运维机器人	可视化大盘	
功能特性	自动根因分析	自动产出解决方案	自动化运维配置	自动分析开放能力	自处理能力	多云联动分析
技术特性	机器人学习	特征库	模式匹配	模型训练	
	实时计算		离线计算		
数据源（租户层、业务层）	操作系统级别数据	用户业务日志数据	用户平台操作数据	云监控数据	历史分析数据

智能诊断平台冬奥实践

冬奥某业务系统突发业务异常应急

①业务异常发生

②算法实时动态检测

③算法多维度根因分析

④NLP以及时序数据实现定界

⑤专家经验自动化根因下钻

⑥时序数据清洗分析，定位根因

⑦提供问题处理建议&解决方案

智能诊断首页，可看见全局信息，快速发现风险

智能诊断系统流程

资源及业务指标整合

明确资源授权及业务各项监控核心指标

数据源接入

自动获取，分析数据来自：
云监控、操作审计、配置审计、应用实时监控服务、用户业务日志数据

拓扑发现

通过数据接入自动绘制资源架构拓扑，可视化展示业务逻辑

智能诊断

根据阿里巴巴多年运维沉淀以及阿里云专家经验，使用不同数据流和监控相关指标，以利用机器学习模型快速生成正常的应用程序模式和行为。

- 阈值分析
- 自动监控
- 根因下钻

聚合分析

使用机器学习算法，将监控项异常与操作事件关联起来，以实时产出或者前瞻性的产出处理建议。

告警及API集成

集成告警配置，相关产品配置项和第三方事件管理服务。

从云上冬奥到大型活动保障

云上大型活动挑战

流量洪峰

高峰流量达到百万QPS级别，需要大量资源预留调度，深度性能优化

全局复杂度

涉及多业务承压运行和调度，带来系统复杂度和内部组织协同复杂度

超大数据量

短时间达到TB级别数据交换和处理，对大数据处理能力要求高

故障零容忍度

大型活动需要确保零故障，对整体稳定性及应急预案要求极高

云上大型活动保障策略

容量评估及调优

通过全链路压测感知系统瓶颈，以容量预置和性能调优进行治理

架构优化及模块化

梳理和优化业务及云平台架构，实现模块化和全局接口可调用性

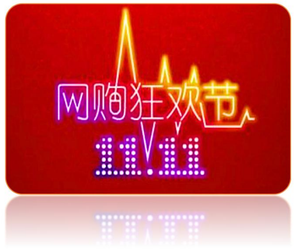
优化数据处理

区分实时处理和离线处理业务，对实时数据处理逻辑做优化

稳定性治理及演练

利用监控和应急体系保障稳定性，活动前完成高可用性方面演练

大型活动案例



目录

- 多角度洞察运维痛点
- 企业演进过程中运维解决之道
- 阿里集团用云最佳实践
- 企业运维平台的未来

运维平台的发展趋势

企业围绕应用、云服务、云平台、基础设施构建可观测、可自动化智能化运维全新的云运营、云工具策略和云运维模式成为必然趋势.

Gartner 2022 IT运维与云管关键趋势指出

2022 key in IT Operations and Cloud Management

1、加快应用程序发布速度将需要新的运营和工具策略

不可变基础设施和基础设施即代码（IaC）作为应用程序操作和工具的核心原则。

2、从基础设施向平台和服务的转变，需要纪律性的自动化，这是云运维的基础

使用DevOps和开发实践使自动化成为IT和业务的弹性基础。称为持续基础架构自动化（CIA）。

3、云增加的复杂性需要新的架构元素和成熟的云运营模式

使用云”Landing Zones”以增强管理和治理的最佳实践

4、对可观察性和效率的需求不断增长会加速APM，DEM和AIOPS的使用

可观察性（observability）是一个属性而不是过程

最小化、持续的投资AIOPS组合来获取即时价值

5、自助服务和去中心化将需要成熟的治理和ITSM实践——在分布和敏捷的时代，IT服务管理并没有消失

使用策略即代码（PaC）来加强安全性和合规性

发展ITSM实践以满足数字业务的需求

- 自动化事件管理
- 通过变更管理平衡速度和风险
- 去中心化的配置管理

想一想，我该如何把这些技术应用在工作实践中？

THANKS

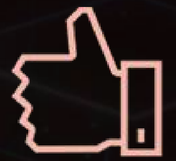
GTS服务介绍

阿里云GTS部门(Global Technical Service) 通过多种服务方案组合帮助您在企业数字化转型全生命周期中实现业务成功

上云/迁云 → 用云/管云

起步

扩张



优化

咨询服务

- Landing Zone
- 云迁移咨询
- 架构设计
- 数仓设计

云迁移及交付服务

- 数据库迁移
- 云原生迁移改造
- 数据中台建设
- 大数据平台建设

企业支持计划

- 运维日常技术支持
- 主动服务

专家增值服务

- 重保护航
- FinOps 解决方案
- 风险评估
- 健康检查
- 容量评估
- 容灾方案

行业领域解决方案

- 电商容灾方案设计
- 安全托管专家服务
- 中国出海网络解决方案
- 培训认证方案

垂直领域解决方案

- 云网络专家服务
- 安全加固专家服务
- 数据类专家服务(数据库/大数据)
- CDN直播点播专家服务