广发证券
稳定性保障体系建设实践

2024年10月

https://sre-elite.com

个人介绍

周健华

资深交易系统SRE

个人简介：

- 从业十余年，一直从事运维相关工作
- 2018年加入广发证券，负责集中交易系统、极速交易系统、条件单系统等多套T0、T1级交易系统运行维护
- 参与及推动公司内多项稳定性保障体系建设工作

目录

CONTENT

SREelite

01

稳定性保障体系建设思考

SREelite

稳定性保障体系建设思考

SREelite

痛点

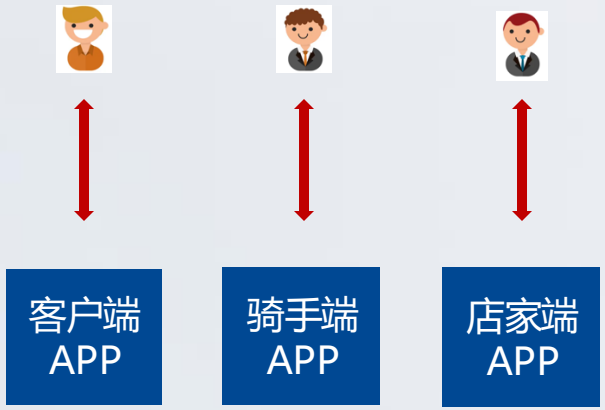
- 系统稳定性要求高
- 故障处理时效要求高
- 外部系统依赖多
- 应用环境复杂

运维过程需要管理的

“人”、“事”、“物”

复杂度高

由一个外卖场景引发的思考



外卖场景		稳定性保障场景
客户下单、店家接单、骑手接单、平台运营、客户反馈、...	场景	变更实施、故障发现、应急协同、故障定界、故障复盘、...
推送规则、时效计算规则、体验反馈规则、...	机制	变更规程、预案规范、评审规则、应急协同流程、...
客户喜好、GIS数据、派送时间、店家评价、客户反馈、...	数据化	执行时间、监控数据、业务数据、复盘数据、...
感知风险、体验分析，持续优化管理机制	分析	感知风险，用户反馈，持续优化场景效率
多端APP、手机GIS定位、连接对象	连接	消费多平台能力、场景与场景间互联、ChatOps快速串联人、事、场

稳定性保障体系建设模型

SREelite

基于场景驱动的平台化管理模式，构建“组织、流程、平台、场景”四位一体的运维场景平台

场景

变更发布	故障应急	OnCall	常规运维	重大保障	周末测试	...
CAB评审	异常申报	值班排班	系统风险评估	关键巡检推送	服务状态公告	...
操作时序	快速集结	盘中巡检	应急预案与演练	业务状态大屏	临时变更	...
自动化发布	应急操作	盘后分析	系统效能	每天回顾	交易日前重点巡检	
变后验证	故障复盘	清算感知	架构韧性评估	告警升级	重点集结	

流程

参考借鉴

ITIL　ISO 20000　DEV OPS　SREelite　Big Data AIOps Machine Learning

外部&公司规范

- 容灾等级
- 信息安全
- 合规风控
- ...

运维规程

- 发布规程
- 监控规范
- 操作规程
- ...

平台

监、管、控、析

Zabbix	CMDB	CD持续交付	运维数据平台	智能算法分析
Prometheus	ELK/日志	脚本平台	RPA自动化	...

组织

技术专家

机器人

Chat Bot

LLM Bot

运维稳定性保障建设全景

SREelite

SRE涉及的活太多，我们挑些来聊聊

02

稳定性保障-事前

SREelite

变更管理

SREelite

变更是稳定性的第一杀手

对于证券公司更是丝毫不敢懈怠，因为变更后业务流量很可能面临上线即高峰的处境

交易系统的变更一般有按周迭代、变后测试验证、首日重点保障的特点

周末测试验证：每一个周末都犹如经历一次系统“大手术”

业务提供服务 or 不提供服务？

使用生产环境 or 灾备环境 or 独立环境？

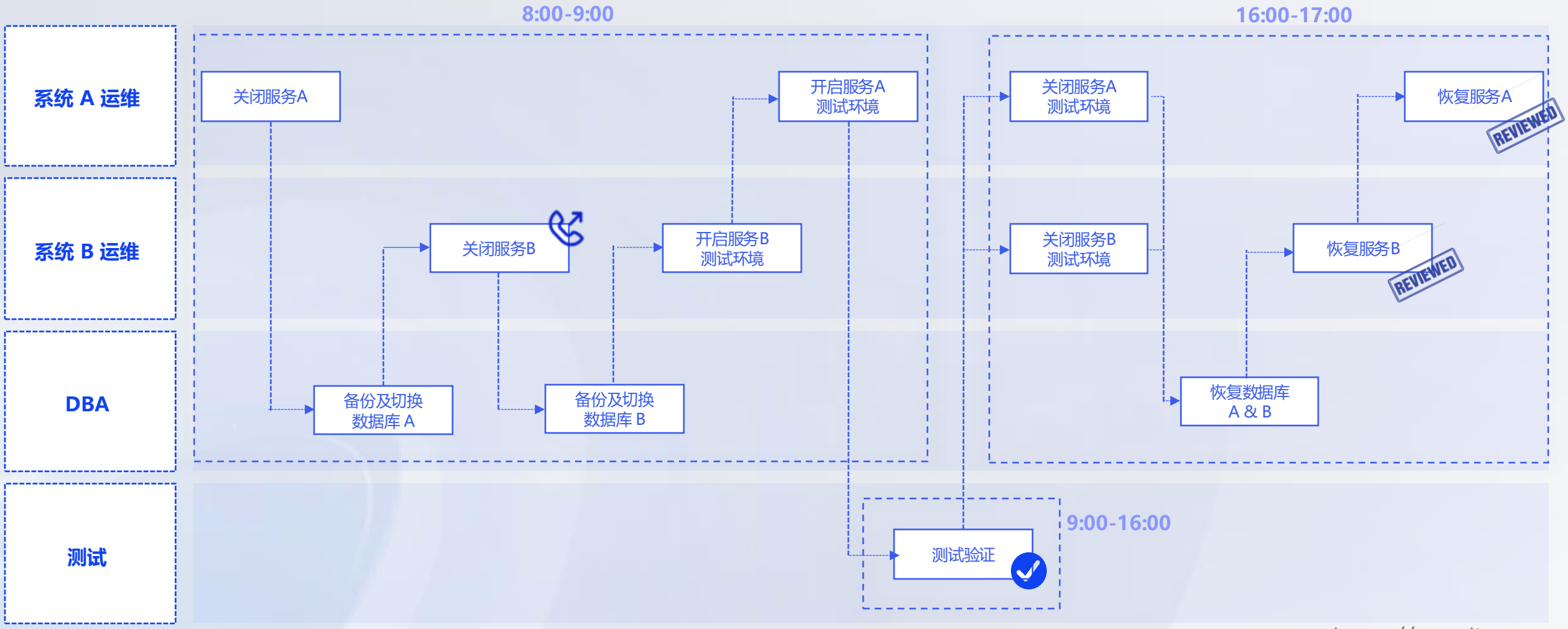
操作风险巨大！

60+重要系统同时参测、数十项测试任务配合、上百人参与测试

变更管理：操作时序工具

SREelite

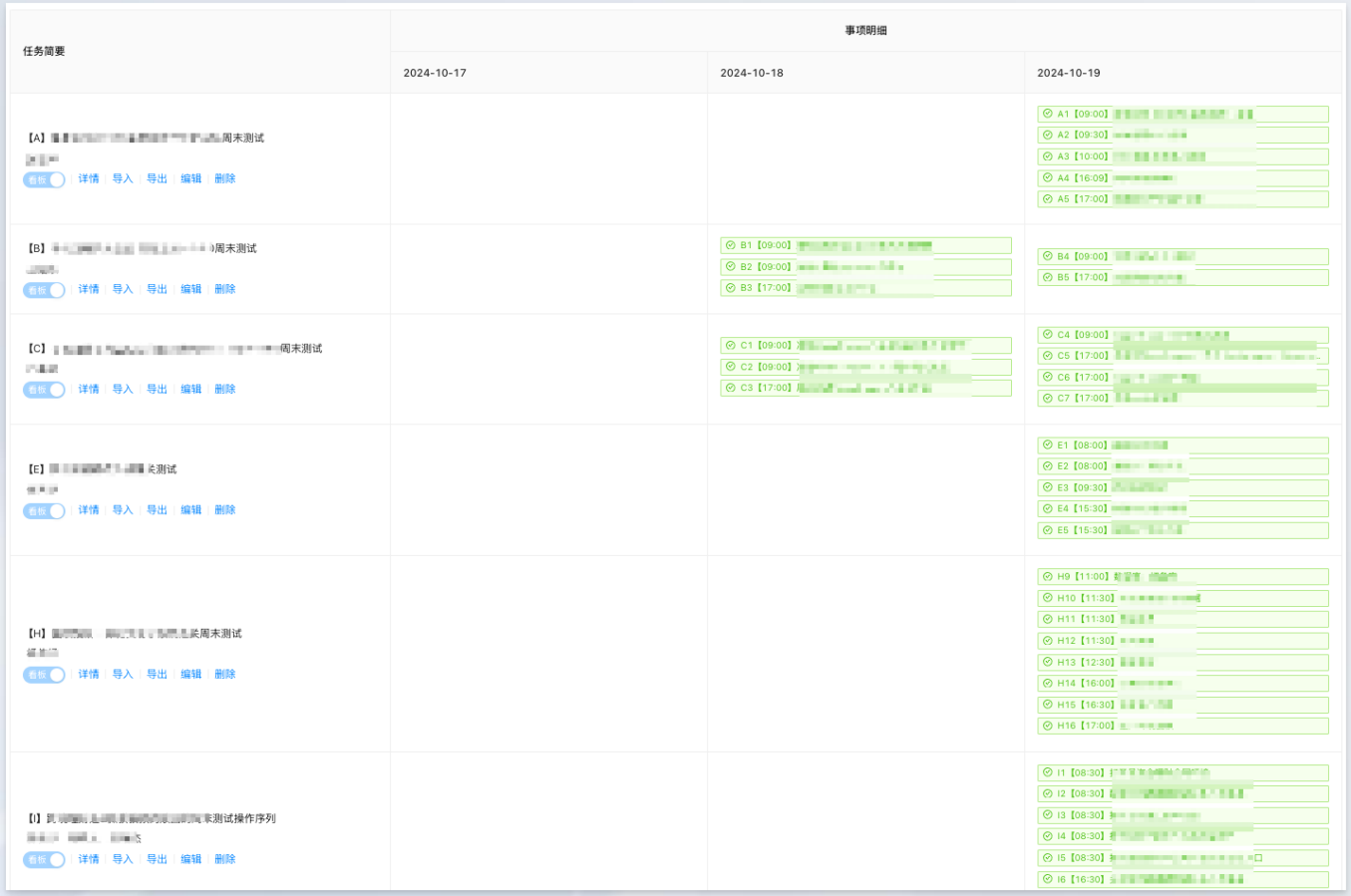
操作时序工具，确保需要各岗位人员间协同的操作不早做、不晚做、不漏做



变更管理：操作时序工具

SREelite

提前沟通对齐信息与导入，时序工具**自动按时提醒**，**逾期自动升级**督办



时序配置汇总

个人任务列表

逾期升级督办

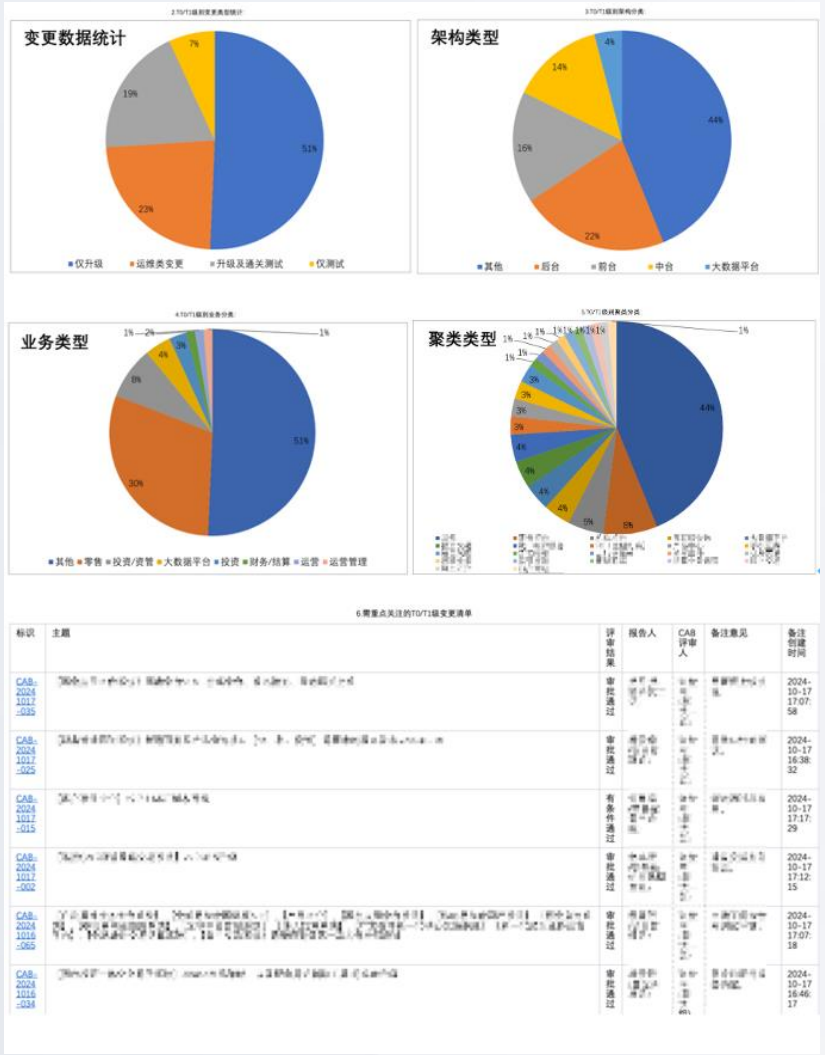
变更管理

SREelite

多个场景工具为变更管理中多个环节开道护航

- 变更日历：直观感知变更数量、周末测试环境类型
- 变更评审报告：重要变更公示，打通各系统间变更信息壁垒

变更日历



变更评审汇总报告

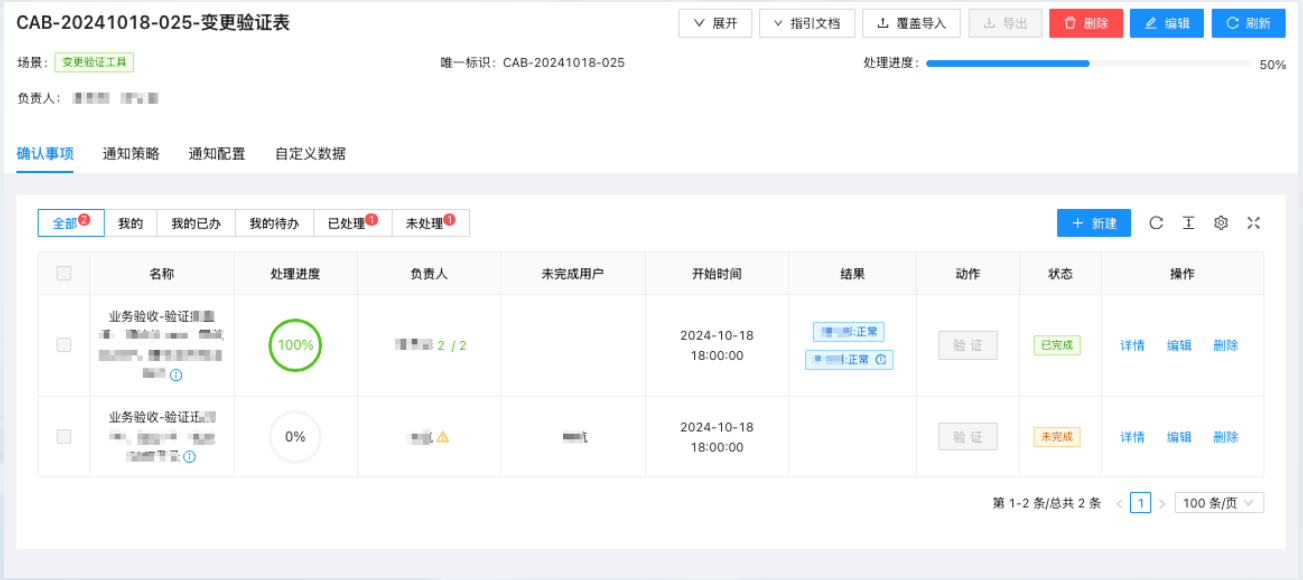
变更管理

SREelite

多个场景工具为变更管理中多个环节开道护航

- 变更验证表：自动变更验证提醒督办，确保开闸前完成验证闭环
- 集结通知：自动化通知相关人员集结保障
- 巡检推送：聚焦关键时刻时系统、数据状态，做到心里有数

集结通知



变更验证表

关键时刻重要巡检推送

应急预案

SREelite

预案线上化、数字化

“像搭乐高一样编写应急预案”

制作可复用的原子预案，将原子预案进行组合编排形成完整预案

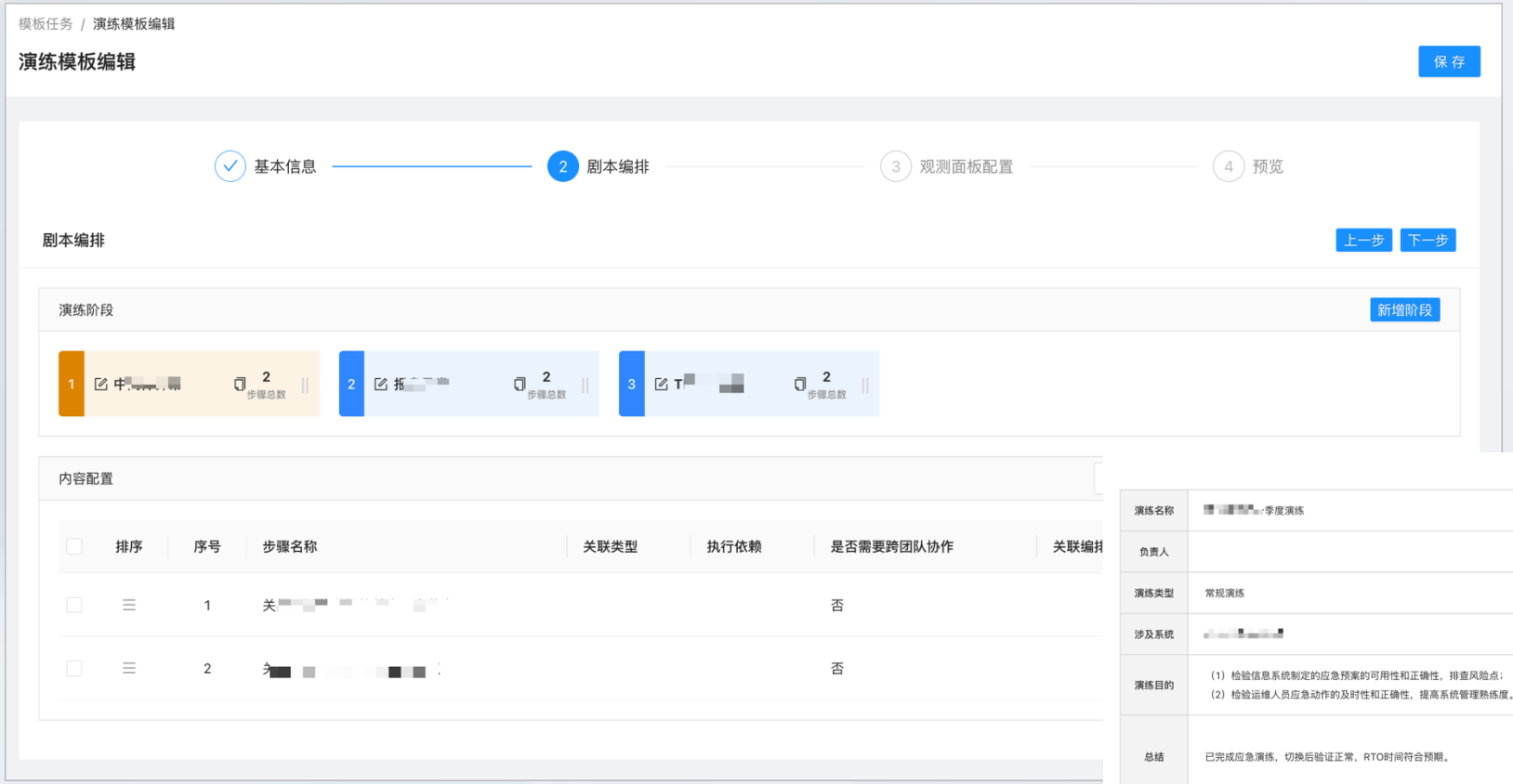
应急演练

SREelite

应急演练的过程是基于应急预案的验证和度量

应急预案与演练

SREelite



季度演练

演练名称	季度演练	演练时间	2024-08-10 11:48:52 - 2024-08-10 12:05:53
负责人		所属群组	
演练类型	常规演练	是否涉及风险	否
涉及系统			
演练目的	（1）检验信息系统制定的应急预案的可用性和正确性，排查风险点； （2）检验运维人员应急动作的及时性和正确性，提高系统管理熟练度。		
总结	已完成应急演练，切换后验证正常，RTO时间符合预期。		

演练方案

演练阶段	序号	步骤	关联编排预案	执行结果	预期执行时长	实际执行时长	与预期时长偏离(倍)	执行人
中间件异常	1	点	干预	已完成	30秒	23秒	↓23%	
	2	点		已完成	1 分钟	1 分钟	↓0	
报盘异常	3	子		已完成	1 分钟	40 秒	↓33.3%	
	4	务		已完成	3 分钟	2 分钟 20 秒	↓22%	

*注:建议根据演练结果调整预计时间

02

稳定性保障-事中

SREelite

应急保障体系

IT运营指挥中心（ECC）

- 主要包括故障监测、应急指挥、数据运营、重大演练等核心职能
- OnCall 人员的值守、应急作战指挥室的线下场所

应急保障体系

SREelite

OnCall 与 应急角色

Google SRE 实践

故障指挥官 Incident Commander

全面协调和管理整个故障响应过程，做出关键决策，确保团队高效协作

通讯官 Communications Lead

负责内外部沟通，及时更新利益相关者，管理事故状态页面，协调跨团队合作

运维指挥 Operations Lead

深入技术分析，执行故障诊断和修复操作，提供专业技术建议

借鉴 & 适配

统筹管理
资源协调
应急指挥
值班质控
盘后复盘

值班经理
统筹管理

一线运维
值班岗
快速恢复

在线监测、巡检
异常申报
应急响应
诊断定位
业务恢复

二三线
协同支撑

ChatOps
人机协同

二线运维专家支持
测试复现与验证
研发代码排查与修改
产品业务逻辑分析
项目资源协调

监控告警响应机器人
应急管理辅助机器人
值班经理助手
OnCall机器人
巡检机器人

应急保障体系

SREelite

应急全流程

想方设法缩短故障修复时间（TTR）

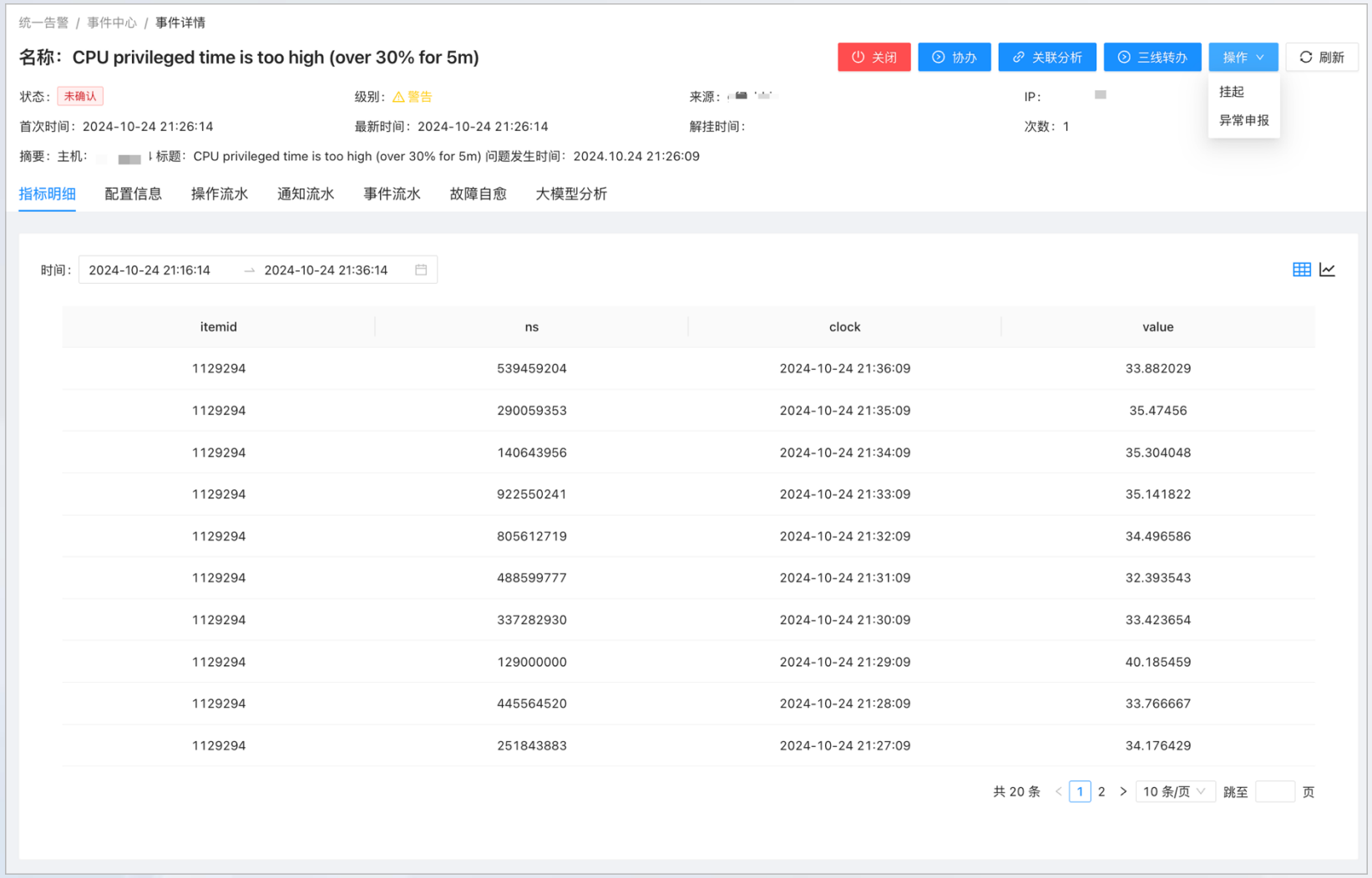
故障预防	故障发现	故障响应	故障定界	故障恢复	复盘改进
风险评估	监控发现	报警触达与升级	应急指挥	应急三把斧	过程复盘
应急演练	主动巡检	异常识别	故障定界	架构韧性	根因分析
监控及感知完善	协同反馈	异常申报	故障观测	紧急维护	周边清查
性能压测	主动拨测	影响分析	异常复现	业务应急	风险揭示
协同机制	智能检测	异常集结	智能排障	恢复验证	问题跟踪
……	……	……	……	……	……

故障发现：统一告警

SREelite

不只是“告警看板”，还是排障、应急的流量入口

故障发现：统一告警


统一告警 / 事件中心 / 事件详情
名称：CPU privileged time is too high (over 30% for 5m)
关闭
协办
关联分析
三线转办
操作
刷新
挂起
异常申报
状态：未确认
级别：警告
来源：
IP：
首次时间：2024-10-24 21:26:14
最新时间：2024-10-24 21:26:14
解挂时间：
次数：1
摘要：主机：
标题：CPU privileged time is too high (over 30% for 5m) 问题发生时间：2024.10.24 21:26:09
指标明细
配置信息
操作流水
通知流水
事件流水
故障自愈
大模型分析
时间：2024-10-24 21:16:14 2024-10-24 21:36:14
itemid ns clock value
1129294 539459204 2024-10-24 21:36:09 33.882029
1129294 290059353 2024-10-24 21:35:09 35.47456
1129294 140643956 2024-10-24 21:34:09 35.304048
1129294 922550241 2024-10-24 21:33:09 35.141822
1129294 805612719 2024-10-24 21:32:09 34.496586
1129294 488599777 2024-10-24 21:31:09 32.393543
1129294 337282930 2024-10-24 21:30:09 33.423654
1129294 129000000 2024-10-24 21:29:09 40.185459
1129294 445564520 2024-10-24 21:28:09 33.766667
1129294 251843883 2024-10-24 21:27:09 34.176429
共 20 条
1
2
10 条/页
跳至
页

故障发现：系统全景看板

SREelite

整合各重要指标，为快速发现异常、排障提供便利，一定程度上解决各监控、告警数据孤岛问题

故障发现与应急响应

SREelite

一线

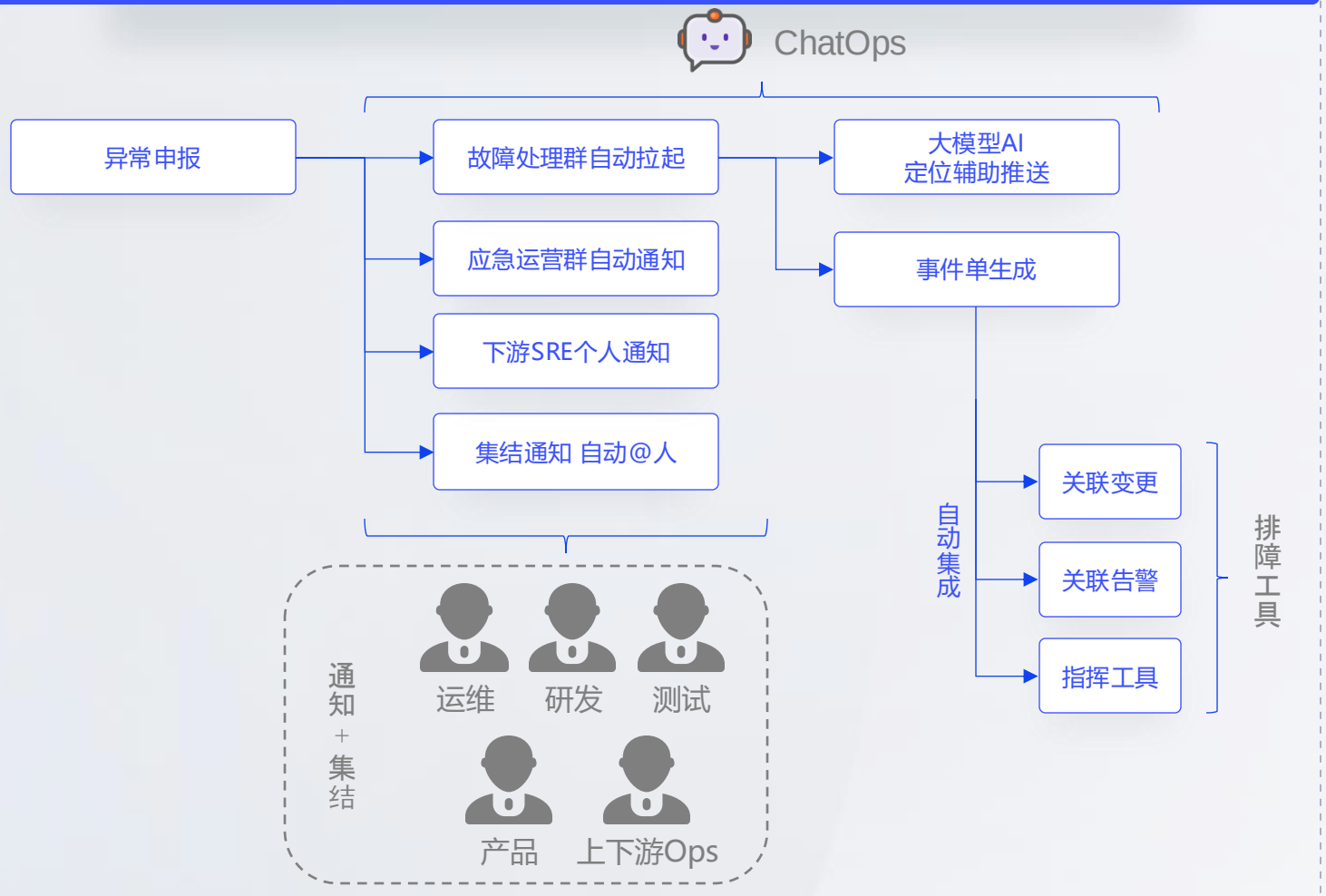
故障发现

- 监控感知
- 巡检发现
- 舆情感知
- 客户反馈
- 情况通报
- 智能预测
- ...

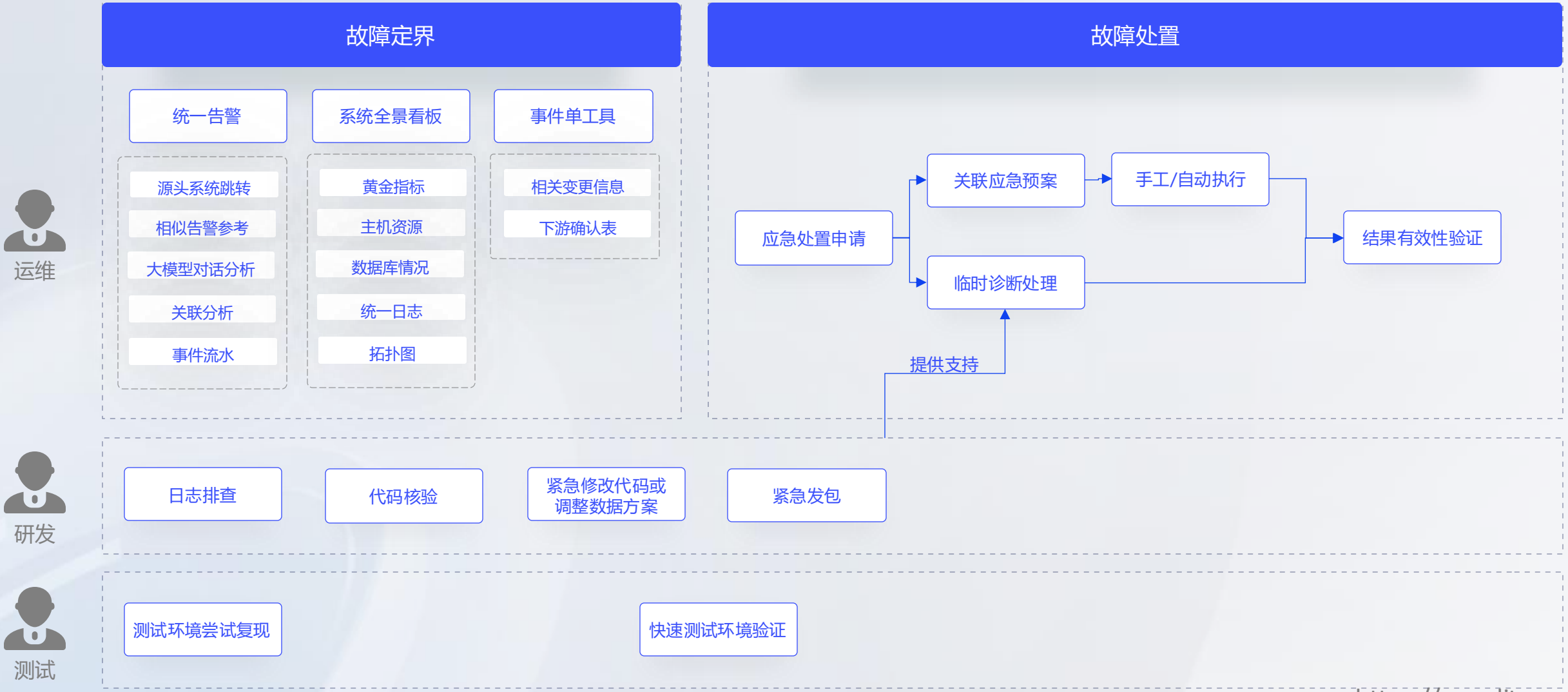
值班经理

介入指挥协调

应急响应

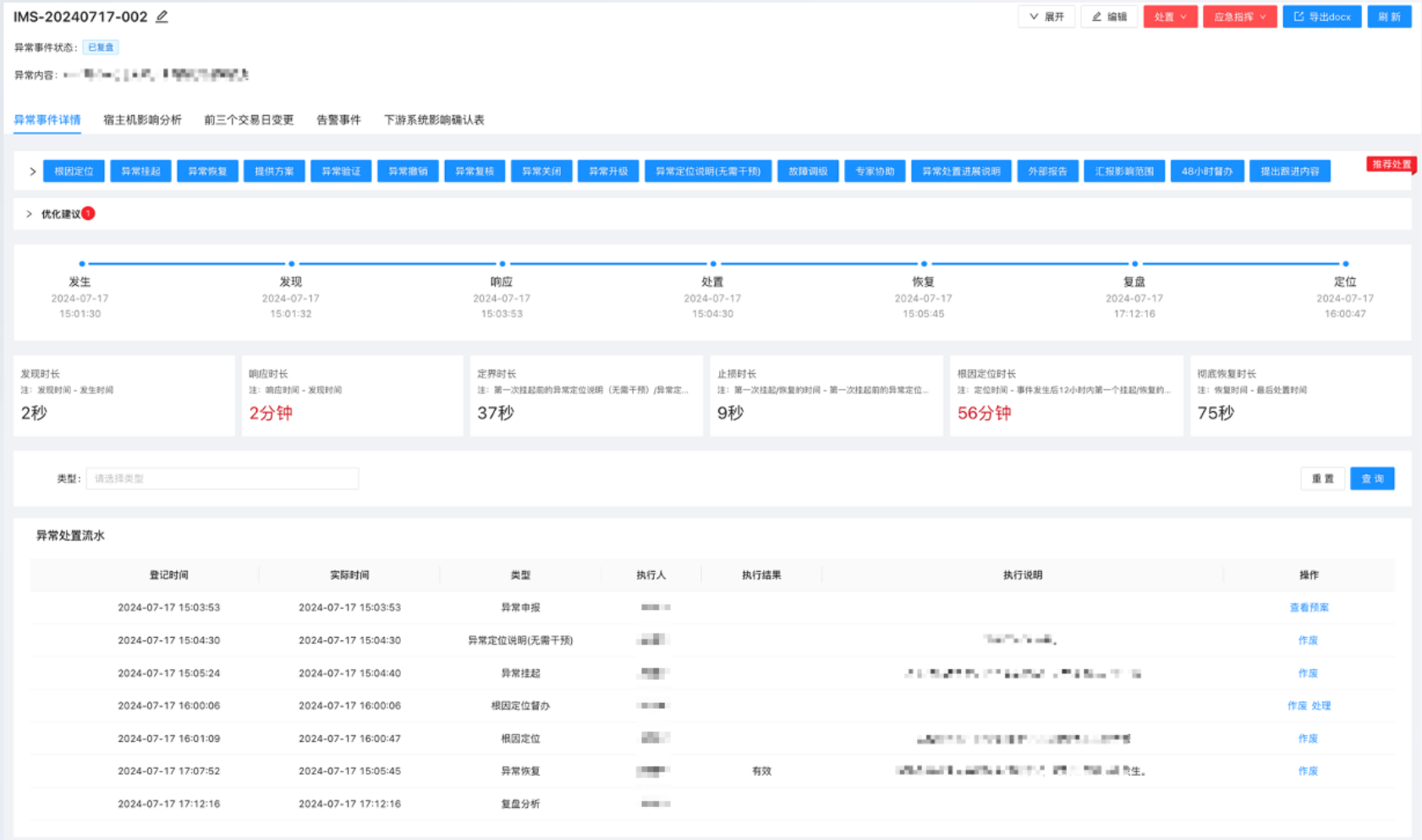


故障定界与故障处置


故障定界
故障处置
统一告警
系统全景看板
事件单工具
源头系统跳转
相似告警参考
大模型对话分析
关联分析
事件流水
黄金指标
主机资源
数据库情况
统一日志
拓扑图
相关变更信息
下游确认表
应急处置申请
关联应急预案
手工/自动执行
临时诊断处理
结果有效性验证
提供支持
运维
日志排查
代码核验
紧急修改代码或
调整数据方案
紧急发包
研发
测试环境尝试复现
快速测试环境验证
测试

标准化事件处理流程

SREelite



02

稳定性保障-事后

SREelite

故障复盘

SREelite

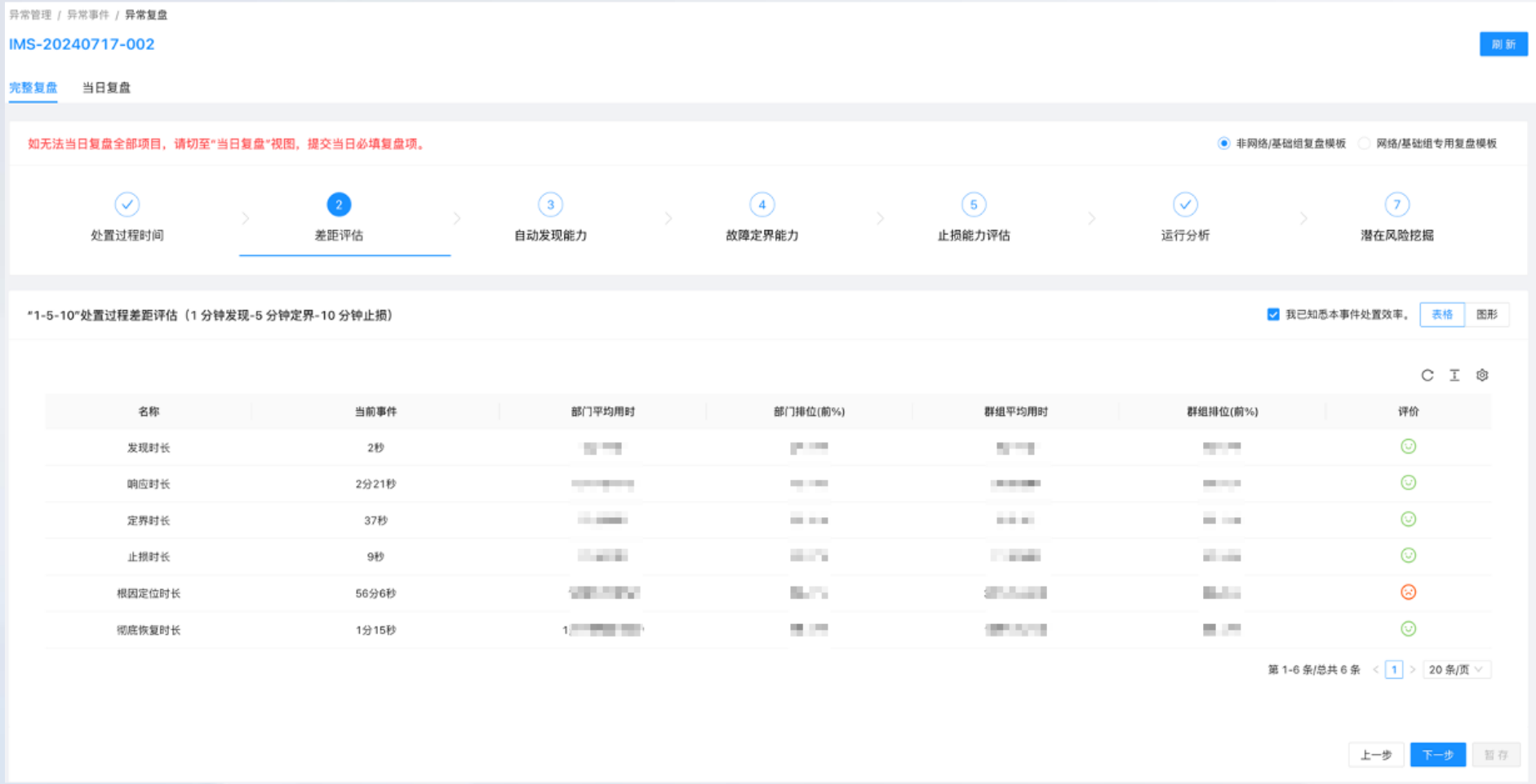
“面向未来”，避免再次踩坑是复盘的最主要目的

每次故障都可能是“事件驱动”改进的机会，通过标准化Checklist方式完整梳理各项能力，查漏补缺

故障复盘

SREelite

标准化复盘CheckList



03

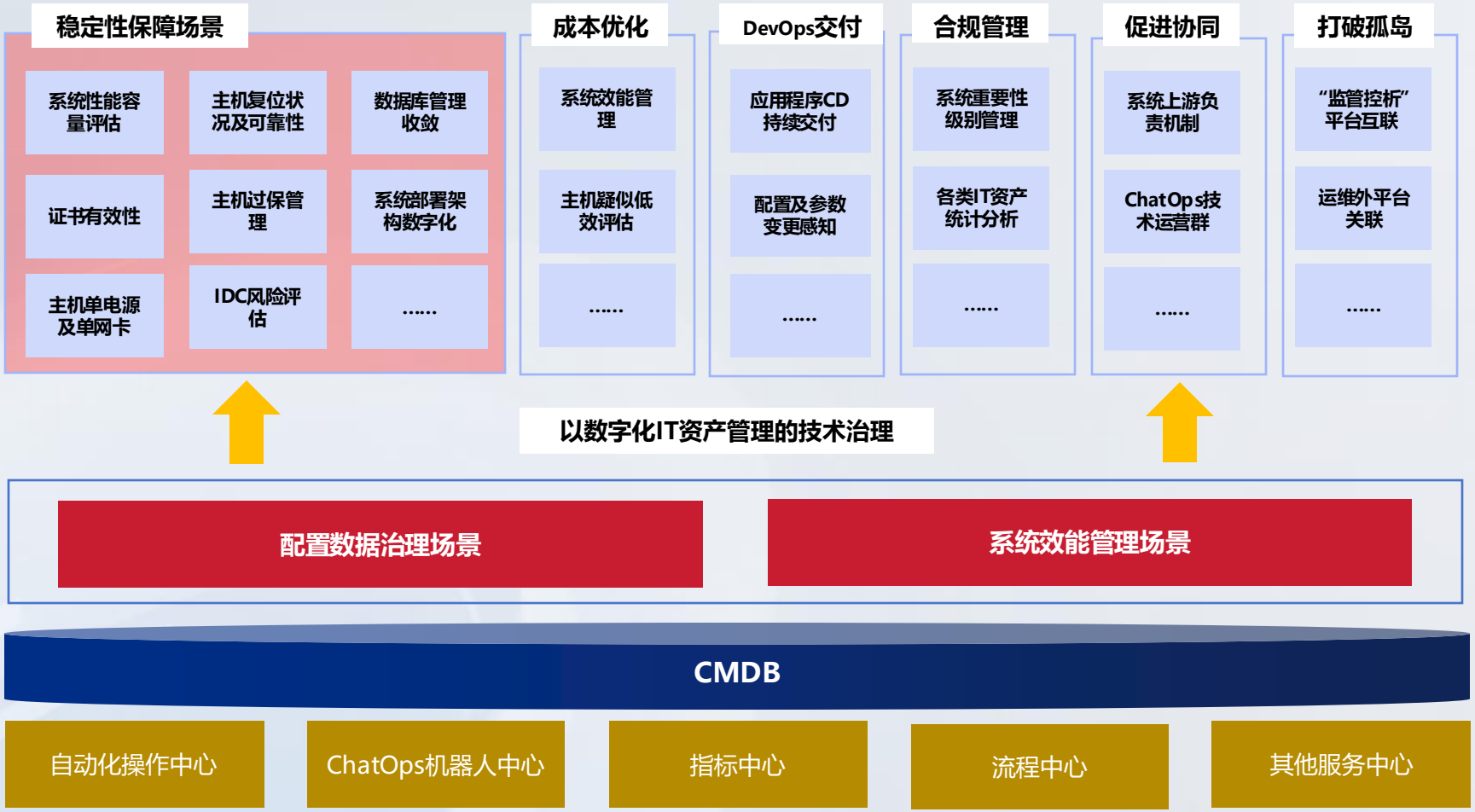
技术治理

SREelite

数据赋能技术治理

SREelite

以数据治理挖掘技术风险，并推动技术治理工作落实



通过对CMDB数据进行挖掘，发现一些技术风险，如：

- 数据库管理收敛
- 证书有效期提醒
- 单电源单网卡
- 单IDC机房风险
- ...

配套数字化看板及专项报告

持续治理及自动提醒

数字化场景的落地与运营

SREelite

“数字化场景上线只是里程碑之一，只有用户使用起来才是场景产生绩效的开始，场景需要专岗技术运营，推动场景真正的解决运维过程中的痛点或工作期望”

每一个**场景来源于运维实际的工作**，沉淀运维专家与管理经验，利用数字化重塑现实的工作
场景保持敏捷的迭代，场景与场景之间形成互联

场景的落地需要配套的规范、规程的支撑，以及数据洞察的支持

每个场景要有需求方、产品经理、研发岗、机器人。其中产品经理尽量来自SRE专家

每个场景的推广有其重要阶段

“四化水平”评估场景数字化程度，驱动场景研发迭代方向

用户团队使用情况查看各团队的“平台化管理”能力水平，在同一平面上驱动各团队管理岗位加强数字化管理

场景分类	场景名称	子类	应用范围标签	场景简述	平台形式	配套流程与规范	配套技术运营看板	涉及机器人	需求方	产品owner	研发owner	上线状态	推广优化级	四化水平				各组使用情况					
														线上化	数字化	自动化	服务化	A组	B组	C组	D组	E组	F组
应急管理	应急演练	发现风险	全局	常态化的生产系统应急演练计划、任务、流程、运营等工作	场景	043号规程	14	演练管理机器人、时序机器人	张三	李四	王五	已发布(迭代少)	中	基本完成	部分完成	部分完成	不涉及	已使用	已使用	已使用	未使用	已使用	已使用
	容量评估	风险管理	全局	建立围绕指标、策略、报告风险评估的主动运行评估的运行评估场景工具	场景	055号规程	36	容量管理机器人、自定义机器人	张三	李四	王五	试运行&需推广	高	基本完成	基本完成	部分完成	不涉及	已使用	已使用	已使用	未使用	未使用	未使用
	应急预案	应急处置	全局	应急预案线上化、最小计算单元应急策略、预案场景编排、自动化操作、预案消费等	场景	应急管理规范	4	应急预案机器人，事件机器人、自定义机器人	张三	李四	王五	持续迭代	中	已完成	已完成	部分完成	基本实现	已使用	已使用	已使用	已使用	已使用	已使用

04

总结 & 未来展望

SREelite

总结 & 未来展望

SREelite

基于场景驱动建立适配我司流程的平台化管理模式，使用数字化方式，重塑运维工作场景。结合数据挖掘、体验反馈不断迭代优化场景。

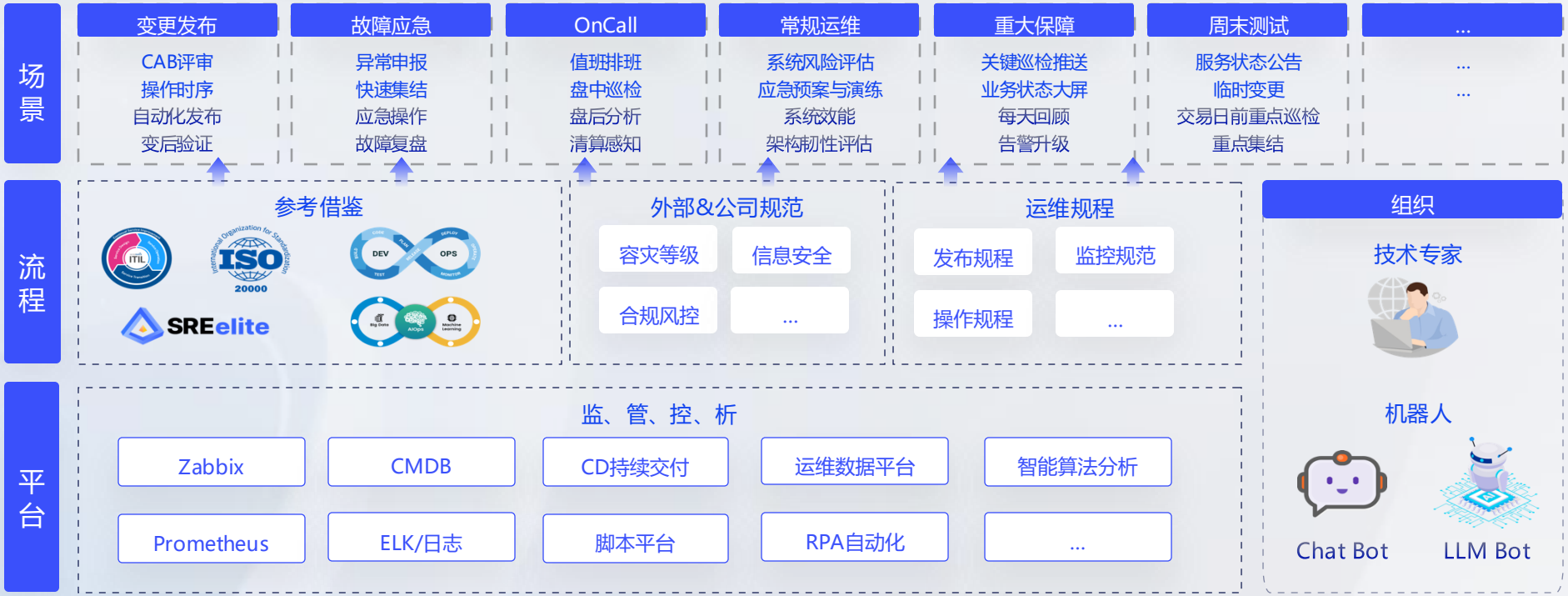
未来展望

流程层：持续借鉴业内先进做法，适配与融合进流程

场景层：持续迭代优化场景，建立更多场景间的连接

平台层：引入可观测性监控体系、全景风险管控平台，加强监、控能力

组织层：深入挖掘LLM Bot的用法，利用AiOps提升效率



Q&A

https://sre-elite.com