

广发证券 稳定性保障体系建设实践

2024年10月



<https://sre-elite.com>



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

01

稳定性保障体系建设思考

02

稳定性保障工作场景

03

技术治理

04

总结 & 未来展望

01

稳定性保障体系建设思考

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

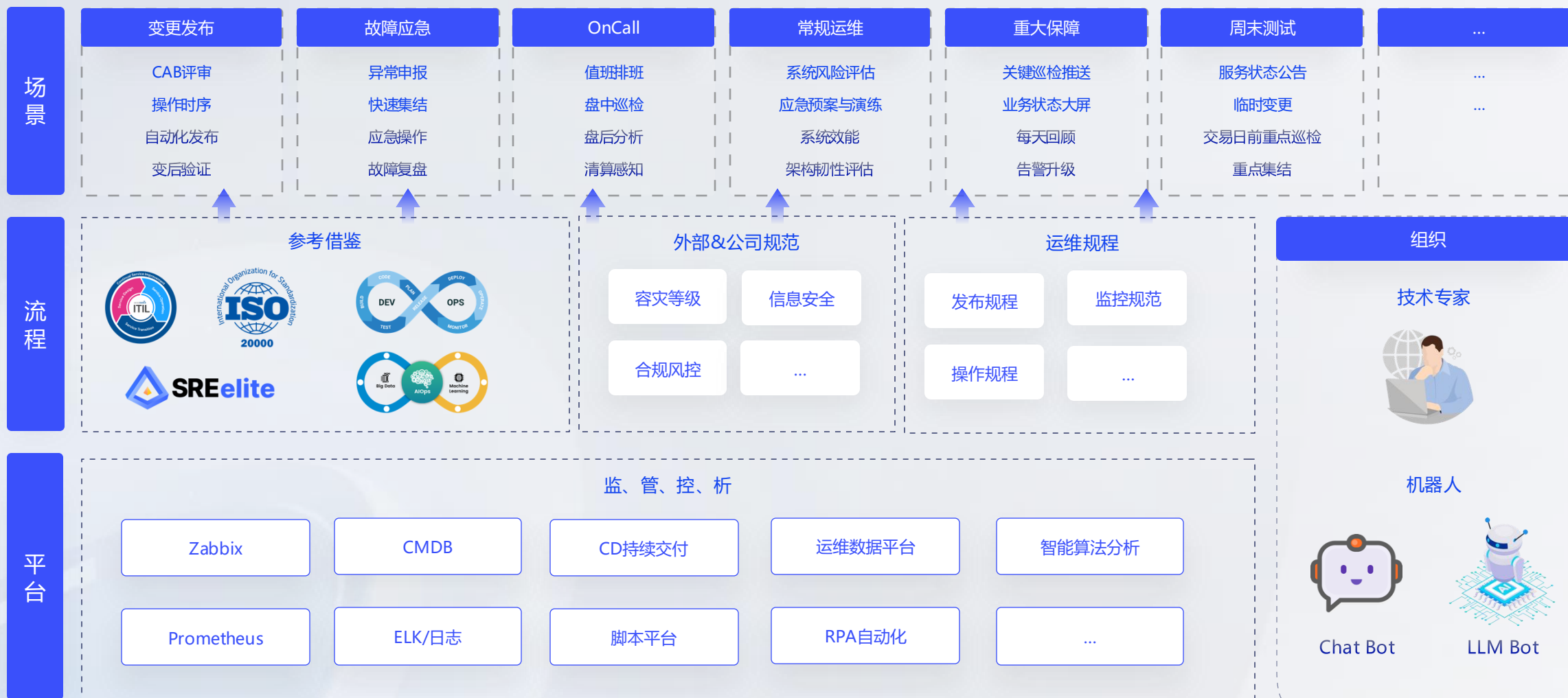


外卖场景

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

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



SRE涉及的活太多，我们挑些来聊聊



02

稳定性保障-事前

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？

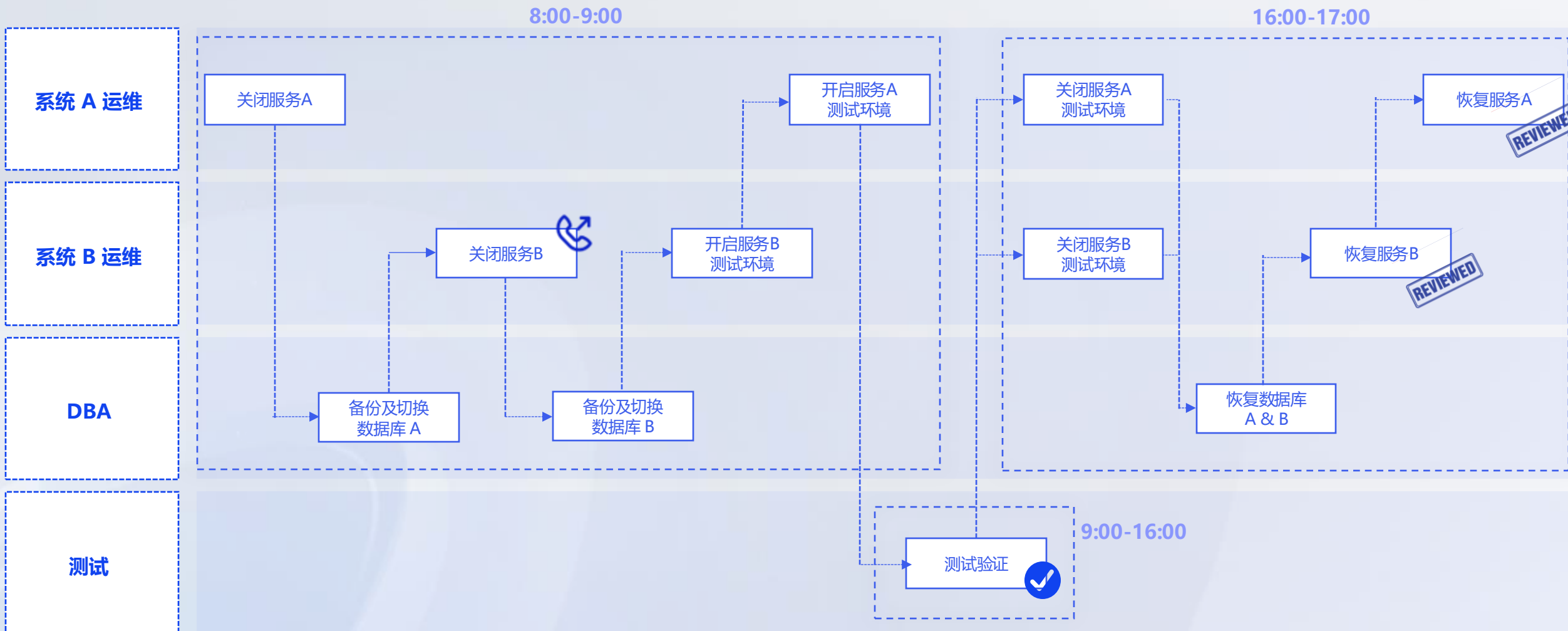
使用生产环境 or 灾备环境 or 独立环境？

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



操作风险巨大！

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**



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**

任务摘要	事项明细		
	2024-10-17	2024-10-18	2024-10-19
[A] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]			<ul style="list-style-type: none">A1 [09:00]A2 [09:30]A3 [10:00]A4 [16:00]A5 [17:00]
[B] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]		<ul style="list-style-type: none">B1 [09:00]B2 [09:00]B3 [17:00]	<ul style="list-style-type: none">B4 [09:00]B5 [17:00]
[C] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]		<ul style="list-style-type: none">C1 [09:00]C2 [09:00]C3 [17:00]	<ul style="list-style-type: none">C4 [09:00]C5 [17:00]C6 [17:00]C7 [17:00]
[E] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]			<ul style="list-style-type: none">E1 [08:00]E2 [08:00]E3 [09:30]E4 [15:30]E5 [15:30]
[H] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]			<ul style="list-style-type: none">H9 [11:00]H10 [11:30]H11 [11:30]H12 [11:30]H13 [12:30]H14 [16:00]H15 [16:30]H16 [17:00]
[I] 周末测试 [按钮: 详情, 导入, 导出, 编辑, 删除]			<ul style="list-style-type: none">I1 [08:30]I2 [08:30]I3 [08:30]I4 [08:30]I5 [08:30]I6 [16:30]

时序配置汇总



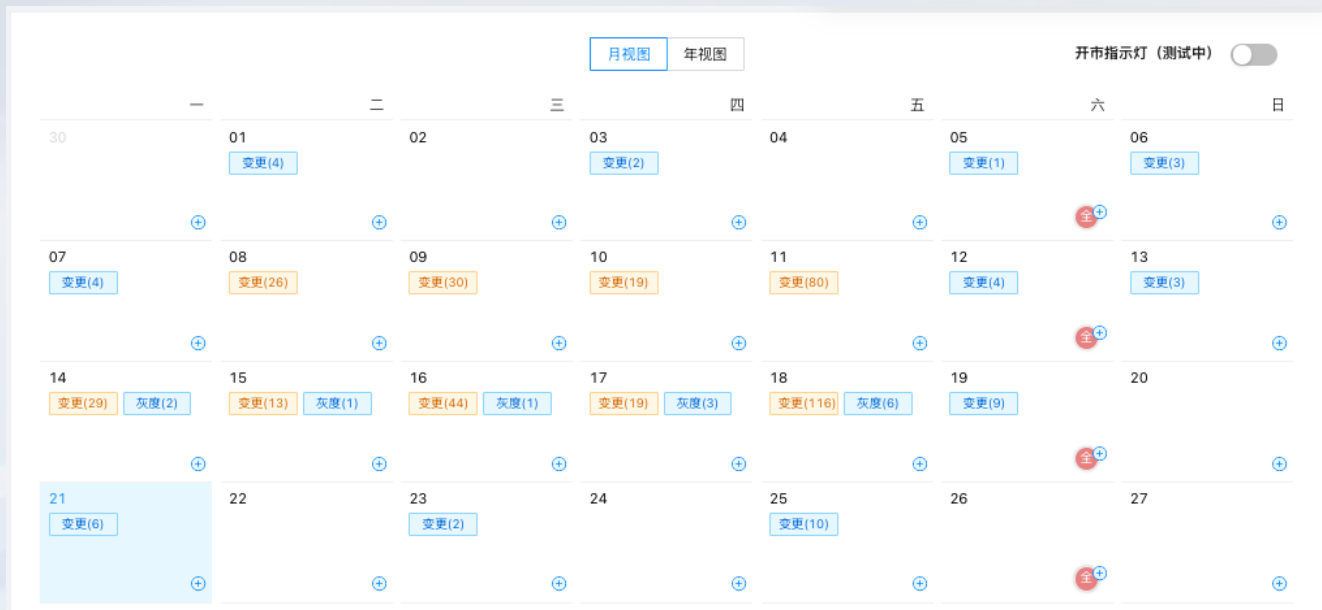
个人任务列表



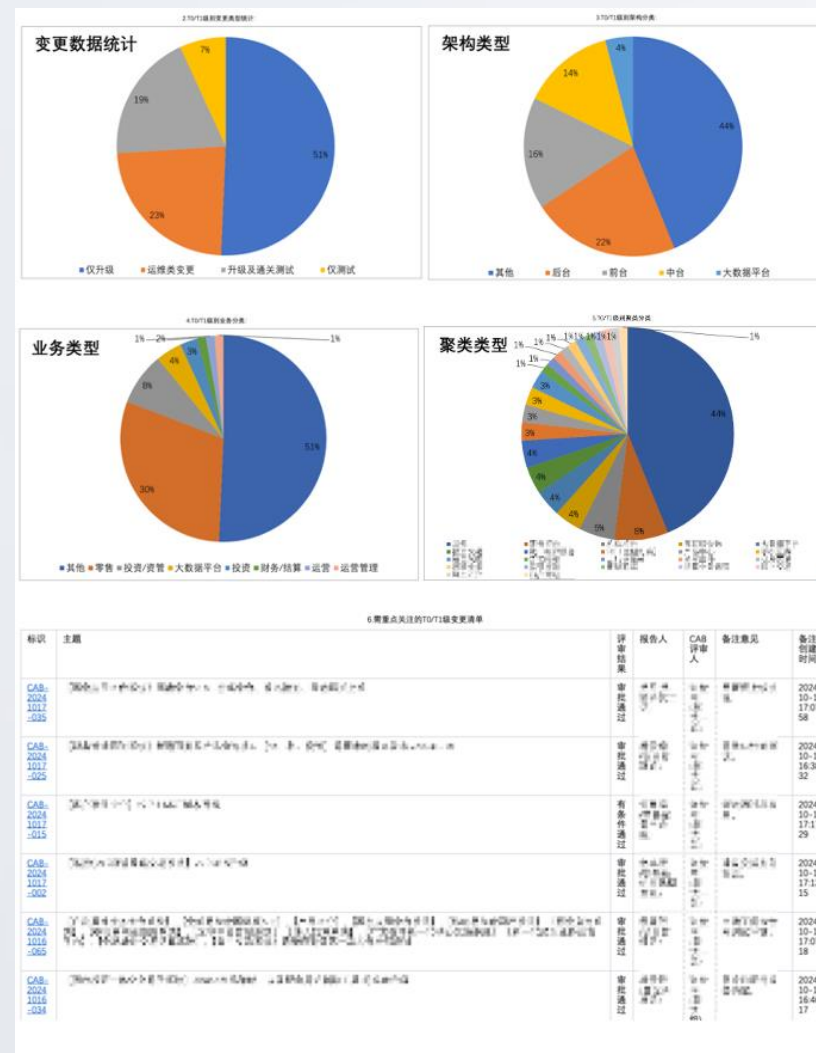
逾期升级督办

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

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

CAB-20241018-025-变更验证表

场景: 变更验证工具 唯一标识: CAB-20241018-025 处理进度: 50%

负责人: [Avatar]

确认事项 通知策略 通知配置 自定义数据

名称	处理进度	负责人	未完成用户	开始时间	结果	动作	状态	操作
业务验收-验证前置	100%	[Avatar] 2 / 2		2024-10-18 18:00:00	验证正常	验证	已完成	详情 编辑 删除
业务验收-验证后置	0%	[Avatar]	[Avatar]	2024-10-18 18:00:00		验证	未完成	详情 编辑 删除

第 1-2 条/总共 2 条 < 1 > 100 条/页

变更验证表



关键时刻重要巡检推送

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



应急演练的过程是基于应急预案的**验证**和**度量**



模板任务 / 演练模板编辑

演练模板编辑

保存

- 1 基本信息
- 2 剧本编排
- 3 观测面板配置
- 4 预览

剧本编排

上一步 下一步

演练阶段

新增阶段

1 中... 2 步骤总数 || 2 拆... 2 步骤总数 || 3 T... 2 步骤总数 ||

内容配置

排序	序号	步骤名称	关联类型	执行依赖	是否需要跨团队协作	关联编排
<input type="checkbox"/>	1	关...			否	
<input type="checkbox"/>	2	关...			否	

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	关...点		已完成	30秒	23秒	↓23%	
	2	关...点		已完成	1分钟	1分钟	↓0	
报盘异常	3	关...点		已完成	1分钟	40秒	↓33.3%	
	4	关...点		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

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



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

二三线
协同支撑

ChatOps
人机协同

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

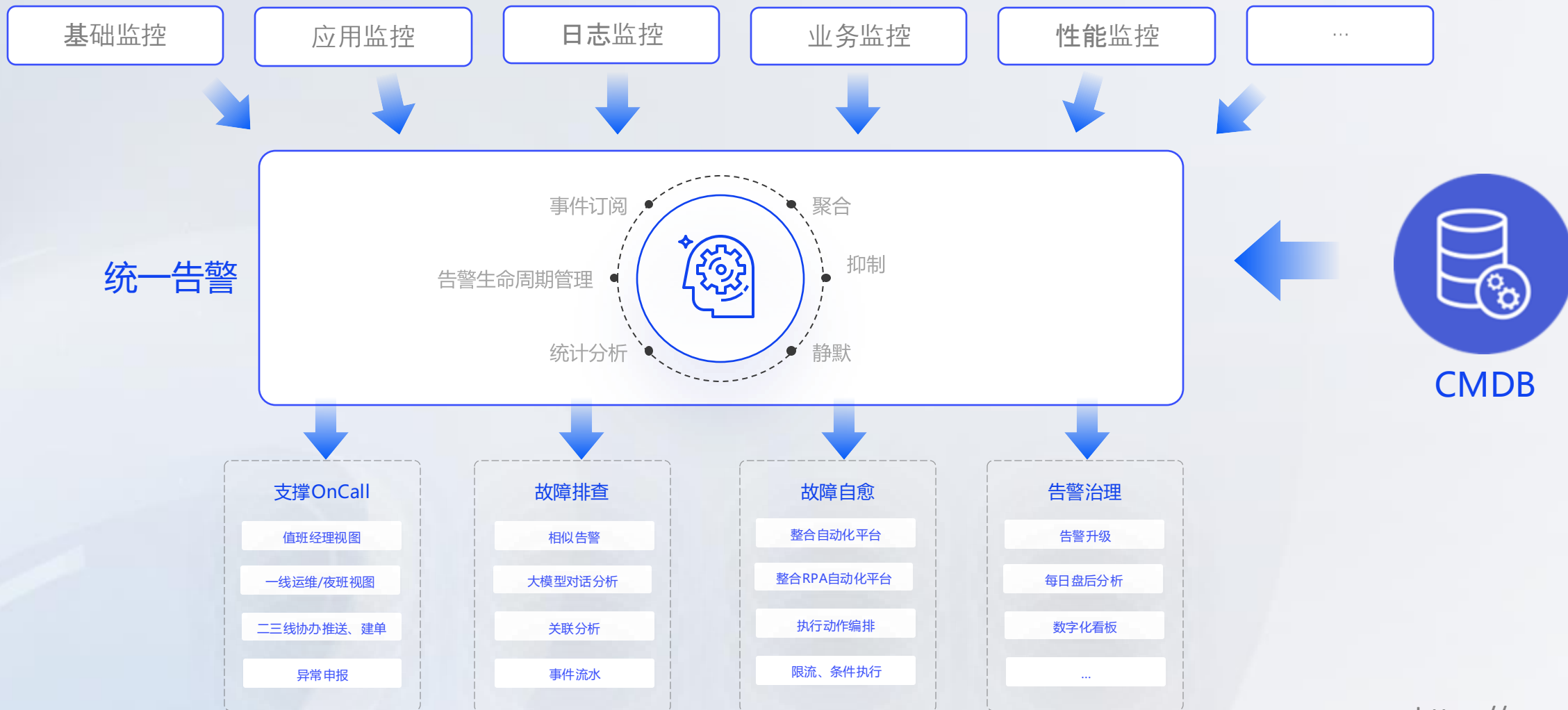
应急全流程

想方设法缩短故障修复时间 (TTR)



故障发现：统一告警

不只是“告警看板”，还是排障、应急的流量入口



统一告警 / 事件中心 / 事件详情

名称： CPU privileged time is too high (over 30% for 5m)

状态：未确认 级别：△警告 来源：[模糊] IP：[模糊] **操作** ▾ 刷新

首次时间：2024-10-24 21:26:14 最新时间：2024-10-24 21:26:14 解挂时间：[模糊] 次数：1

摘要：主机：[模糊] | 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 [网格] [刷新]

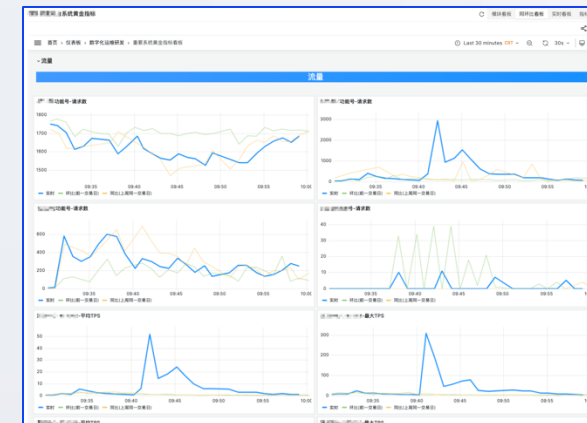
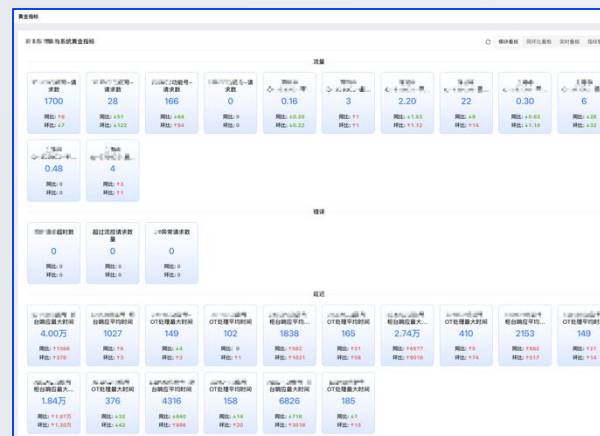
itemid	ns	clock	value
1129294	539459204	2024-10-24 21:36:09	33.882029
1129294	290059353	2024-10-24 21:35:09	35.47456
1129294	140643956	2024-10-24 21:34:09	35.304048
1129294	922550241	2024-10-24 21:33:09	35.141822
1129294	805612719	2024-10-24 21:32:09	34.496586
1129294	488599777	2024-10-24 21:31:09	32.393543
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



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

监控感知

巡检发现

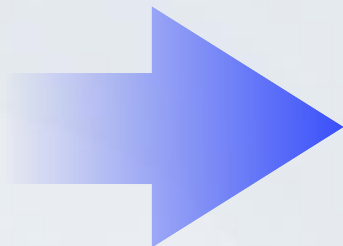
舆情感知

客户反馈

情况通报

智能预测

...



应急响应



异常申报

故障处理群自动拉起

应急运营群自动通知

下游SRE个人通知

集结通知 自动@人

大模型AI
定位辅助推送

事件单生成

通知 + 集结



自动集成

关联变更

关联告警

指挥工具

排障工具



故障定界

统一告警

系统全景看板

事件单工具

源头系统跳转

黄金指标

相关变更信息

相似告警参考

主机资源

下游确认表

大模型对话分析

数据库情况

关联分析

统一日志

事件流水

拓扑图

故障处置

应急处置申请

关联应急预案

手工/自动执行

结果有效性验证

临时诊断处理

提供支持



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



测试环境尝试复现

快速测试环境验证

IMS-2024-07-17-002 | 异常协同

信息技术部运维中心

【异常申报】

注：当前系统本年第 1 次，本月第 1 次异常，点击查看历史事件 (仅内网PC端)

- 1、IMS_ID: IMS-2024-07-17-002
- 2、发生时点: 待定
- 3、发现时点: 2024-07-17 15:01:30
- 4、发生系统: IMS-2024-07-17-002-T1
- 5、发现机制: 监控发现
- 6、异常内容: 异常内容
- 7、是否影响业务: 否, 潜在风险未恢复
- 8、业务影响说明: 潜在风险对系统无影响/已经应急完成, 业务已恢复正常
- 9、容错机制及效果: 具备容错机制并生效
- 10、跟进人员: 跟进人员
- 11、申报人: 申报人

[查看详情](#)

信息技术部运维中心 10-21 8:51

大模型智能定位辅助

事件编号: IMS-2024-07-17-002
发生系统: 发生系统

前三个交易日变更情况:

- 1、CAB-202410-001

应急中心可观测感知 (异常项有才列)

系统健康度: 70

[点击查看](#)

统一告警事件

近2小时告警数: 23

[\[点击查看\]\(https://...\)](#)

BusinessId=5206&StartTime=2024-07-17 06:51:23&EndTime=2024-07-17 08:51:23)

历史相似事件

- 1、IMS-20210-001

推荐应急策略

- 1、IMS-2024-07-17-002 【人工处置】

(注: 数据由大模型筛选生成, 使用时请注意核对)

IMS-20240717-002

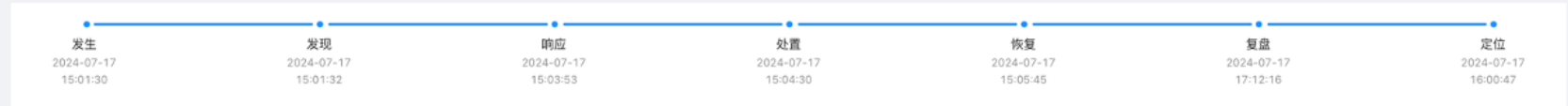
异常事件状态: 已修复

异常内容: 异常内容

[异常事件详情](#) [宿主机影响分析](#) [前三个交易日变更](#) [告警事件](#) [下游系统影响确认表](#)

[根因定位](#) [异常挂起](#) [异常恢复](#) [提供方案](#) [异常验证](#) [异常撤销](#) [异常复核](#) [异常关闭](#) [异常升级](#) [异常定位说明\(无需干预\)](#) [故障调级](#) [专家协助](#) [异常处置进展说明](#) [外部报告](#) [汇报影响范围](#) [48小时督办](#) [提出跟进内容](#) [推荐处置](#)

[优化建议](#)



发现时长 注: 发现时间 - 发生时间 2秒	响应时长 注: 响应时间 - 发现时间 2分钟	定界时长 注: 第一次挂起的异常定位说明 (无需干预)/异常定... 37秒	止损时长 注: 第一次挂起/恢复的时间 - 第一次挂起的异常定位... 9秒	根因定位时长 注: 定位时间 - 事件发生后12小时内第一个挂起/恢复的... 56分钟	彻底恢复时长 注: 恢复时间 - 最后处置时间 75秒
------------------------------	-------------------------------	--	--	--	-----------------------------------

类型: [重置](#) [查询](#)

异常处置流水

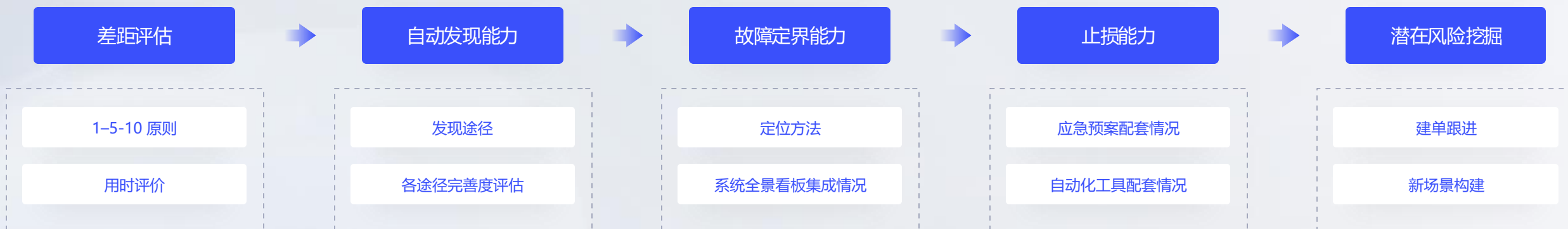
登记时间	实际时间	类型	执行人	执行结果	执行说明	操作
2024-07-17 15:03:53	2024-07-17 15:03:53	异常申报				查看预案
2024-07-17 15:04:30	2024-07-17 15:04:30	异常定位说明(无需干预)				作废
2024-07-17 15:05:24	2024-07-17 15:04:40	异常挂起				作废
2024-07-17 16:00:06	2024-07-17 16:00:06	根因定位督办				作废/处理
2024-07-17 16:01:09	2024-07-17 16:00:47	根因定位				作废
2024-07-17 17:07:52	2024-07-17 15:05:45	异常恢复		有效		作废
2024-07-17 17:12:16	2024-07-17 17:12:16	复盘分析				

02

稳定性保障-事后

“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**



标准化复盘CheckList

异常管理 / 异常事件 / 异常复盘

IMS-20240717-002 刷新

完整复盘 当日复盘

如无法当日复盘全部项目，请切至“当日复盘”视图，提交当日必填复盘项。

非网络/基础组复盘模板 网络/基础组专用复盘模板

处置过程时间 > **差距评估** > 自动发现能力 > 故障定界能力 > 止损能力评估 > 运行分析 > 潜在风险挖掘

*1-5-10*处置过程差距评估 (1分钟发现-5分钟定界-10分钟止损) 我已知悉本事件处置效率。 表格 图形

名称	当前事件	部门平均用时	部门排位(前%)	群组平均用时	群组排位(前%)	评价
发现时长	2秒	00:00:00	100%	00:00:00	100%	🟢
响应时长	2分21秒	00:02:21	100%	00:02:21	100%	🟢
定界时长	37秒	00:00:37	100%	00:00:37	100%	🟢
止损时长	9秒	00:00:09	100%	00:00:09	100%	🟢
根因定位时长	56分6秒	00:56:06	100%	00:56:06	100%	🔴
彻底恢复时长	1分15秒	00:01:15	100%	00:01:15	100%	🟢

第 1-6 条/总共 6 条 < 1 > 20 条/页

上一步 下一步 留存

03

技术治理

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

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**。

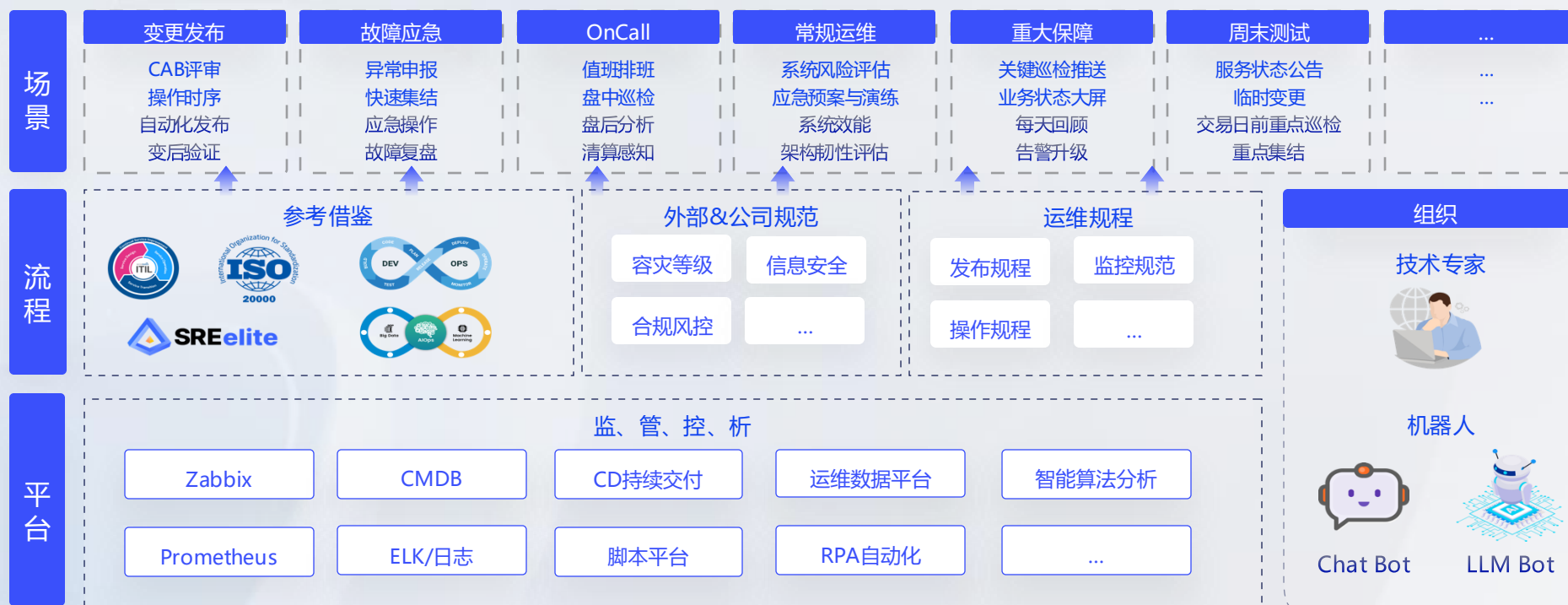
未来展望

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>