中國計算機學會 CHINA COMPUTER FEDERATION　Nightingale　GitLink 确实·开源

第二届CCF·夜莺开发者创新论坛

中国北京　2024.7.26

主办方：中国计算机学会 | 承办方：CCF开源发展委员会、夜莺项目开源社区

像 Google SRE 一样 OnCall - Flashduty 方法与工具剖析

于双羽

快猫星云，Flashduty 产品研发

中国北京　2024.7.26

主办方：中国计算机学会 | 承办方：CCF开源发展委员会、夜莺项目开源社区

大纲

1 **Google SRE 如何 OnCall**

探讨 Google SRE 团队的 OnCall 实践和方法。

2 **借助 Flashduty 降低运维负载**

介绍 Flashduty 如何帮助减轻运维团队的工作负担。

3 **借助 Flashduty 加速告警响应**

讲解 Flashduty 如何提高告警响应速度和效率。

4 **Flashduty 产品路线**

展望 Flashduty 的未来发展方向和潜在改进。

大厂 OnCall 体验比拼

前蚂蚁金服员工[2] ID 6

当年去度蜜月全程背着15寸的电脑，以便随时响应oncall，登机起飞前的几分钟还被拉了个钉钉会议。回来后干了两件事，第一就是申请把电脑换成13寸的，15寸真的重；第二开始刷题换工作。

Amazon员工 ID

oncall到怀疑人生

百度员工 ID

晚上接一个oncall电话，起来处理十几分钟，然后就睡不着了，第二天一天都废了，每次都这样，感觉身体受不了了

Google员工 ID

按小时计费！

字节跳动员工 ID

真想请假写一天代码，每天被oncall搞死

Facebook员工 ID

要oncall，破事儿还贼多

微软员工 ID 4

回复带带da师兄：就这样啊，微软没有运维，所有运维都是工程师自己弄的，很多组都有724oncall值班

蚂蚁金服员工[3] ID 14

现在理解什么是干电池被用尽了吧，我出去旅游什么都可以不带，但是电脑一定得带，在景区的时候都在应急。钉钉一响，整个人就跟惊弓之鸟一样

Google员工[3] ID 2

有的组专门招sre来oncall，但近几年sre的预算卡的紧，很多新的service都是swe在oncall。不过一般oncall都没什么事，还给双倍工资。。所以大家都抢着oncall 😂

《Google SRE 运维解密》

软件工程思维

通过软件工程思维解决运维问题，将技术创新应用于日常运维。

关注研发工作

长期关注研发工作，将琐事占比控制在 50% 以下，保持团队创新力。

琐事过多的影响

- 团队生产力与创新力下降
- 员工士气低落，人才流失
- 系统可靠性与稳定性下降

如何消除琐事？

中断性工作（一般告警）

- 基于 SLO 配置告警
- 告警降噪，将根本原因一致的告警聚合为故障

OnCall（紧急告警）

- 实施值班轮换，同时负责一般告警的处理
- 一个月 OnCall 时间不超过 25%

Google OnCall 方法

文化

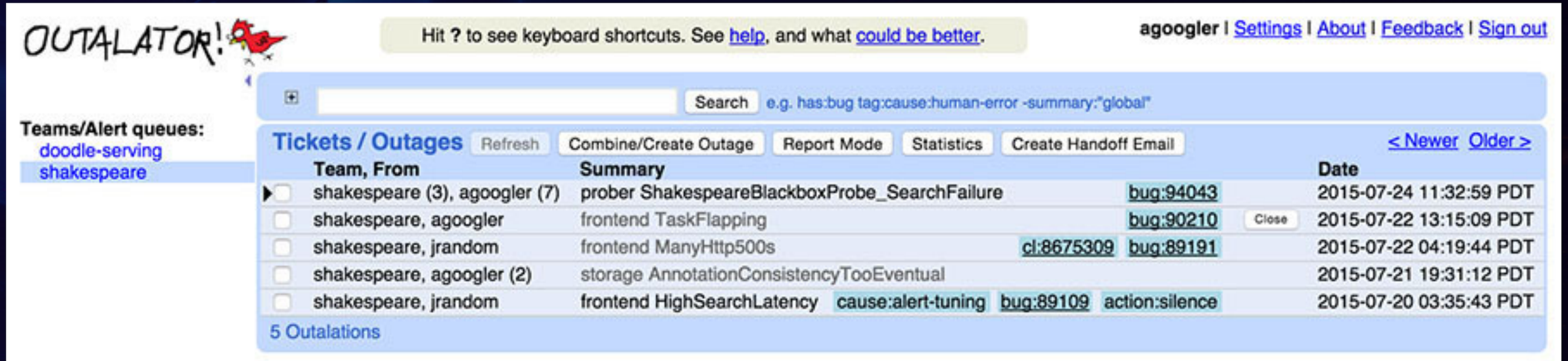
- 强调 OnCall 工作的平衡
- 对 OnCall 工程师进行补贴
- 鼓励事后总结与分享
- 无指责，对事不对人的氛围

机制

- 建立主备 OnCall 值班
- 及时对问题进行响应（SLO）
- 建立清晰的问题升级路线
- 建立明确的故障处置步骤

工具 - Outalator

- 接收公司所有告警
- 将告警降噪为故障
- 对故障打标，数据分析
- 生成故障报告



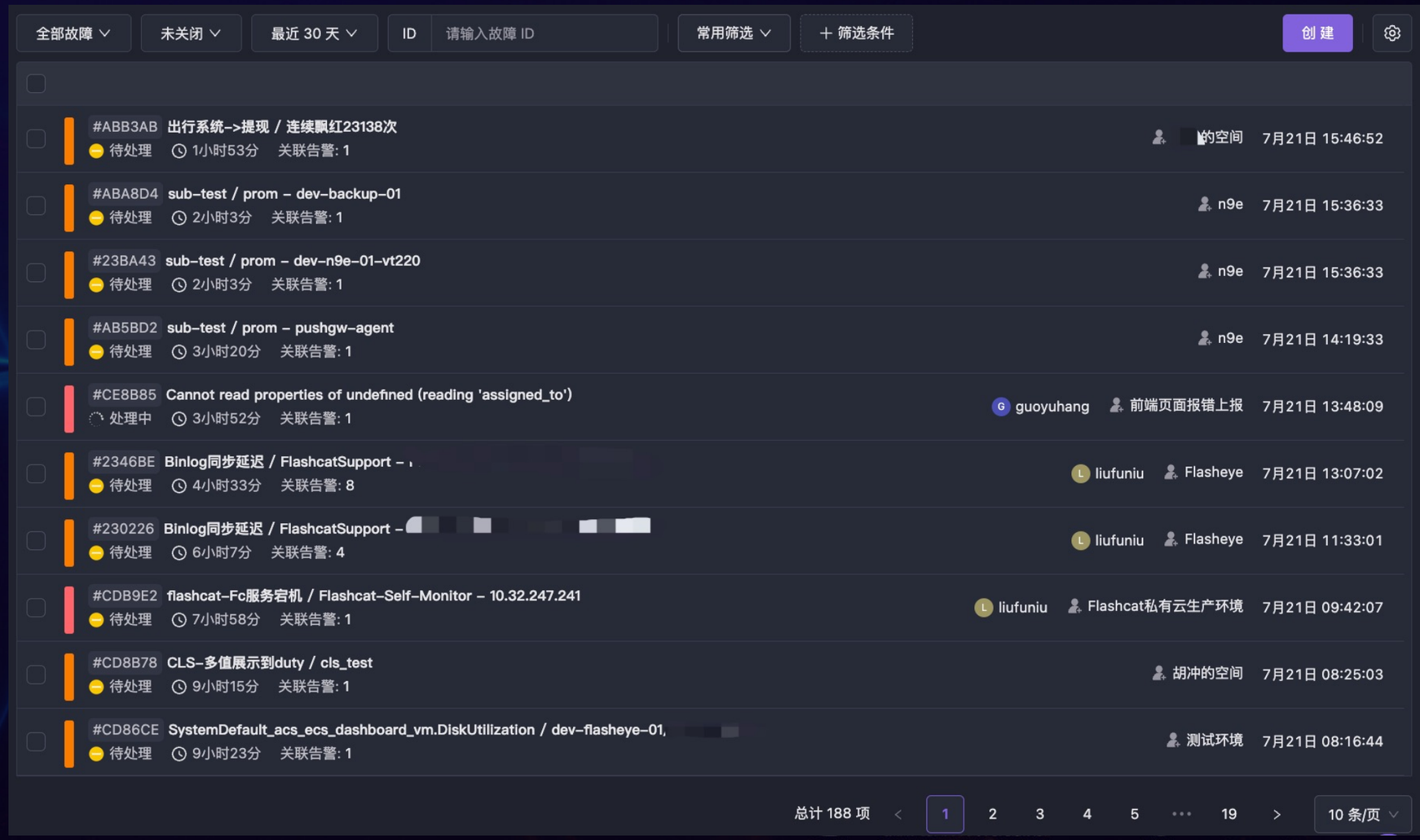
Flashduty，你的Outalator！

降低运维负载

- 对告警进行降噪
- 建立值班机制
- 数据分析，减少长期中断

提升 OnCall 效率

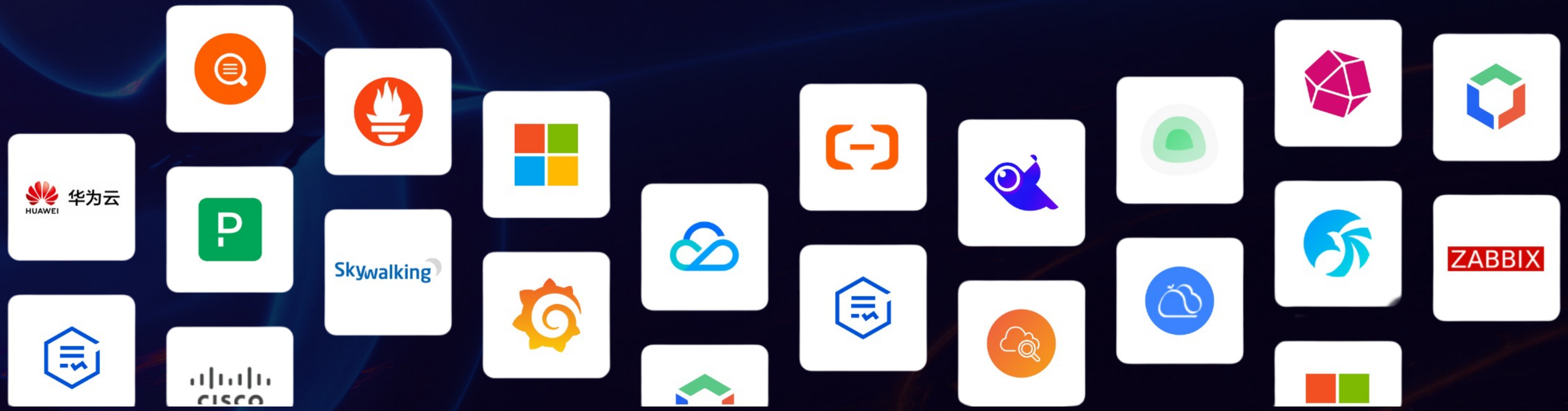
- 明确的故障升级路线
- 故障信息丰富
- IM协同增强
- 自定义操作扩展



中国北京 2024.7.26

集成，一个平台处理所有告警

30+常见监控工具（开箱即用）+ 1 套标准 HTTP 协议 + 邮件解析（覆盖自研监控）



降噪，显著降低告警数量

对相似告警进行聚合

- 事件 => L1 告警 => L2 故障
- 减少通知，避免告警风暴
- 至多降噪 99%

对频发告警进行收敛

- 避免狼来了效应
- 避免频繁被打断

值班，避免整个团队被中断

1 满足各类场景

日常、节假日、调班、限时、公平轮换

2 建立主备值班机制

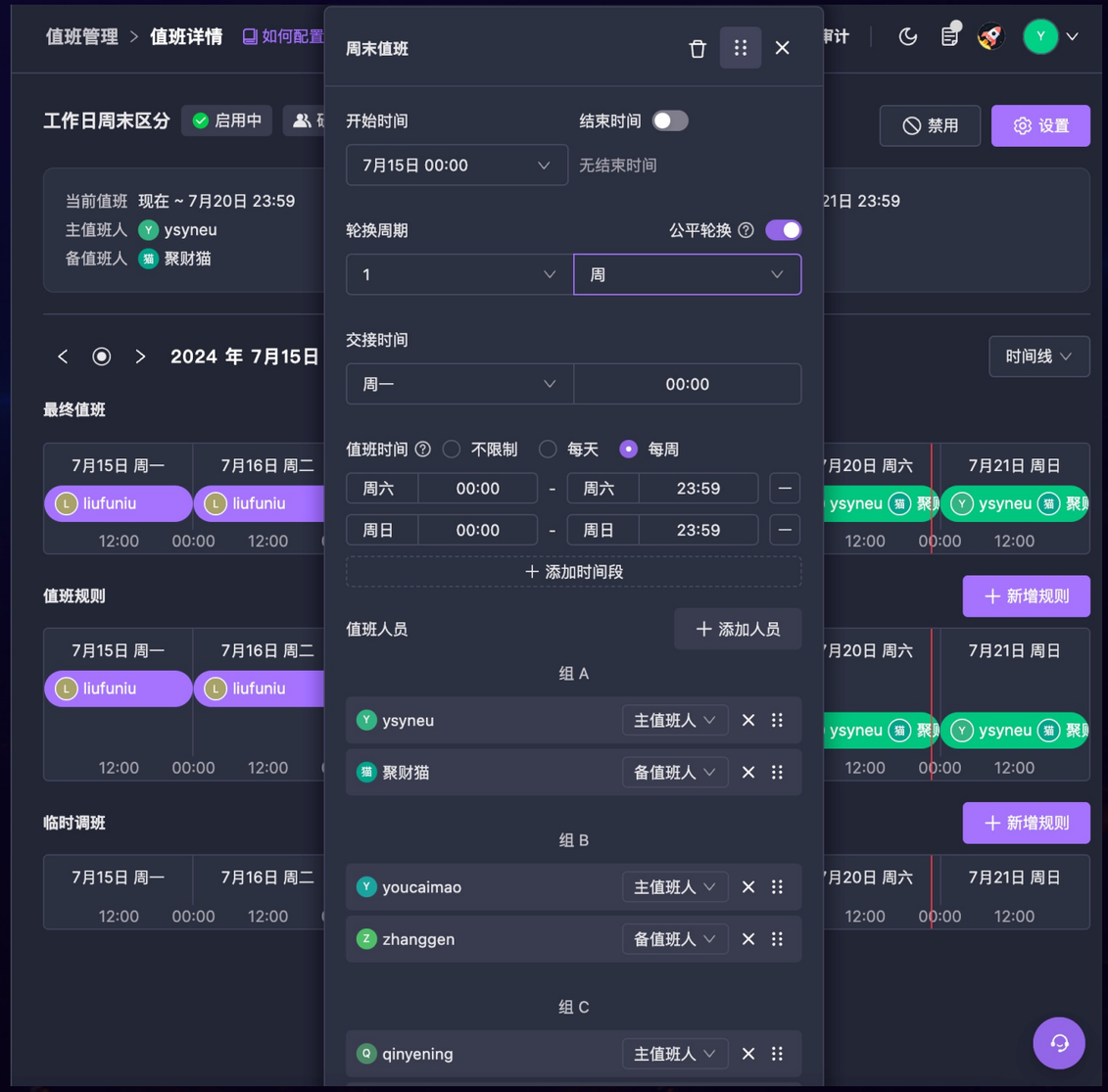
支持多人同时，按角色值班

3 分派告警到值班人

不要随机分派，更不要分派到整个团队

保持住流状态：如果你的目标只是做中断性工作，那么中断性工作就不再是中断。

中国北京 2024.7.26



数据分析，驱动长期改进

运维负载

本周处理了多少告警？周告警数量处于什么趋势？

TopK 告警

哪些告警频繁产生？哪些主机频繁告警？

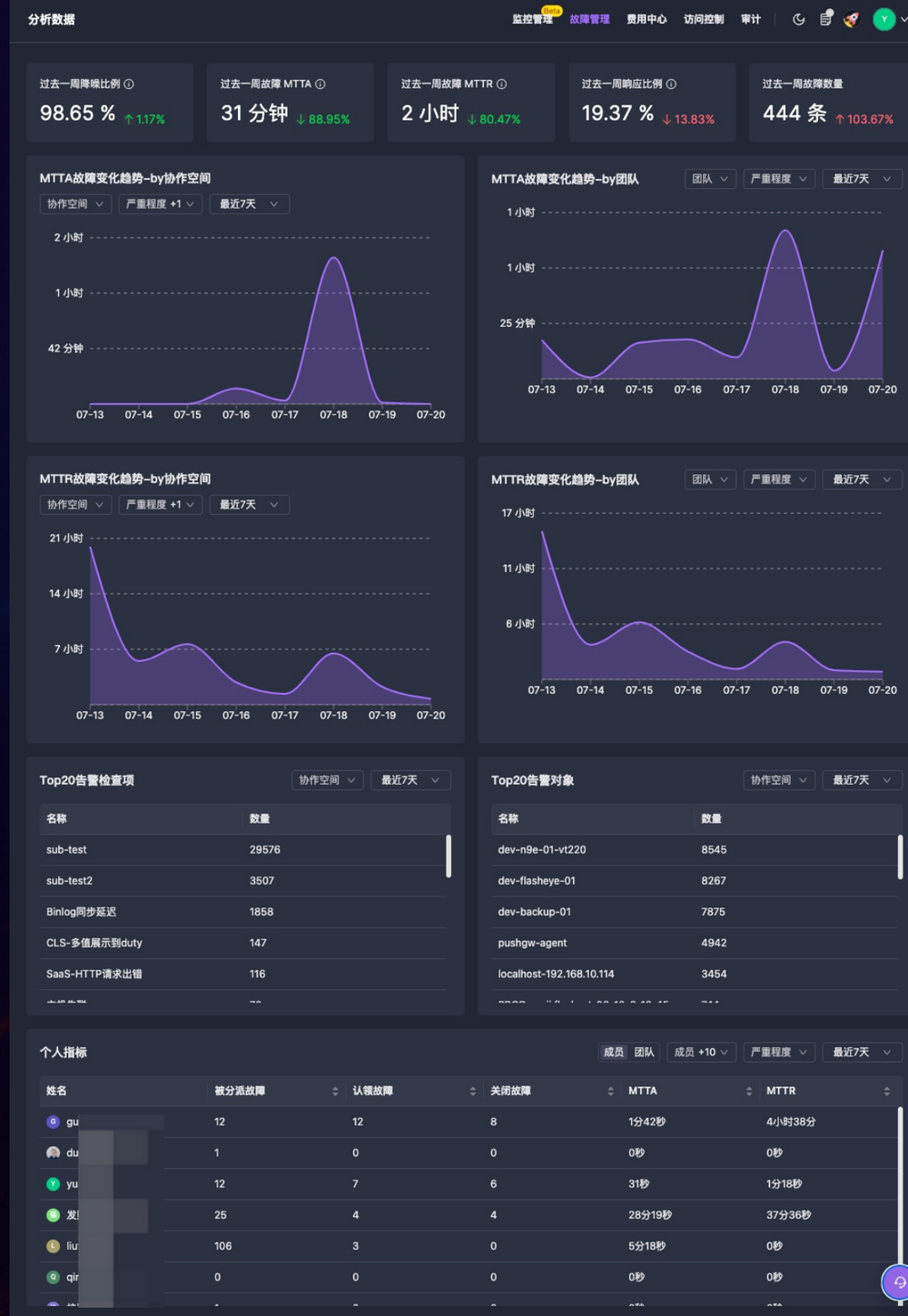
SLO指标

紧急告警的 MTTA 是多少？是否满足 SLO？

可操作性分析

团队对告警的响应比和降噪比如何？

中国北京 2024.7.26



升级，正确时间通知正确的人

1. **制定升级路线**

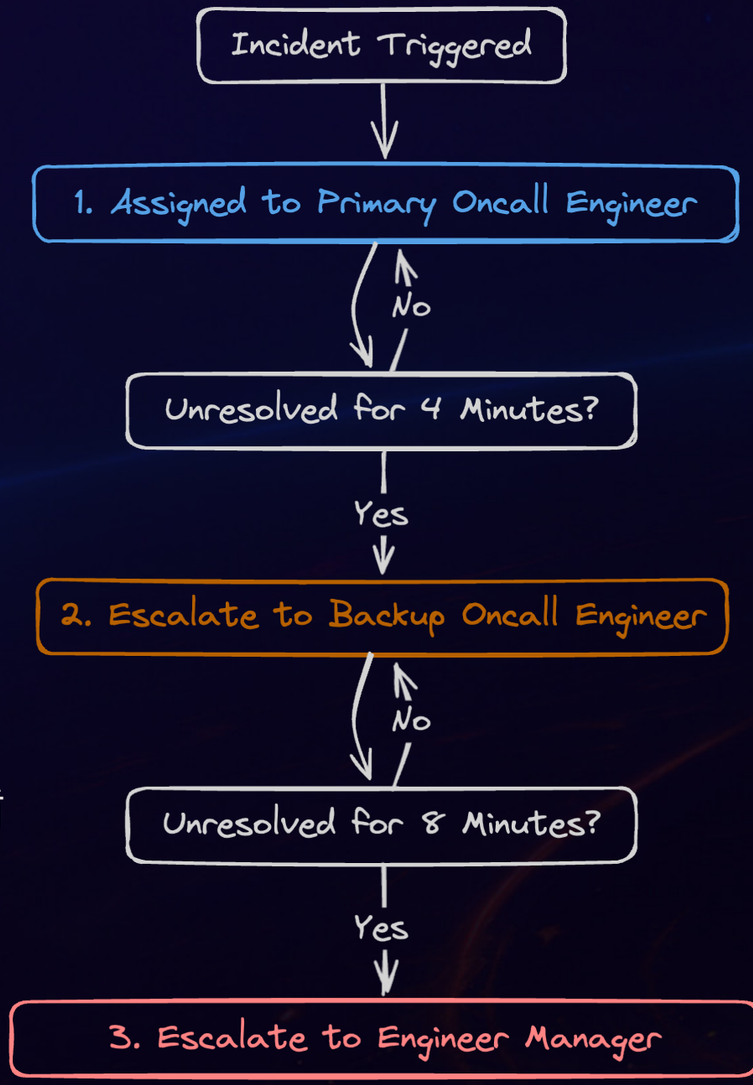
   为故障制定清晰的升级路线

2. **A：确保问题得到响应**

   不被漏处理导致更大资损

3. **B：从容应对紧急情况**

   发生紧急情况不慌张，有后盾



数据增强，丰富故障上下文

历史变更事件

70% 的故障由变更导致

CMDB 元数据

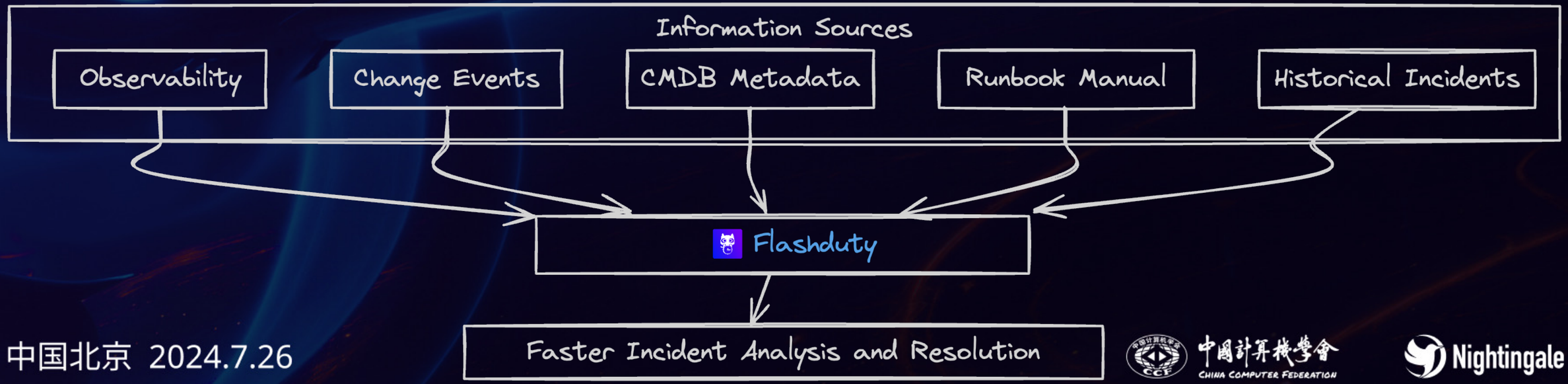
资产关系依赖映射

知识库和 SOP

在故障信息中展示 SOP

历史故障记录

参考相似故障的解决办法



IM集成，随时随地处理故障

1 **实时通知**

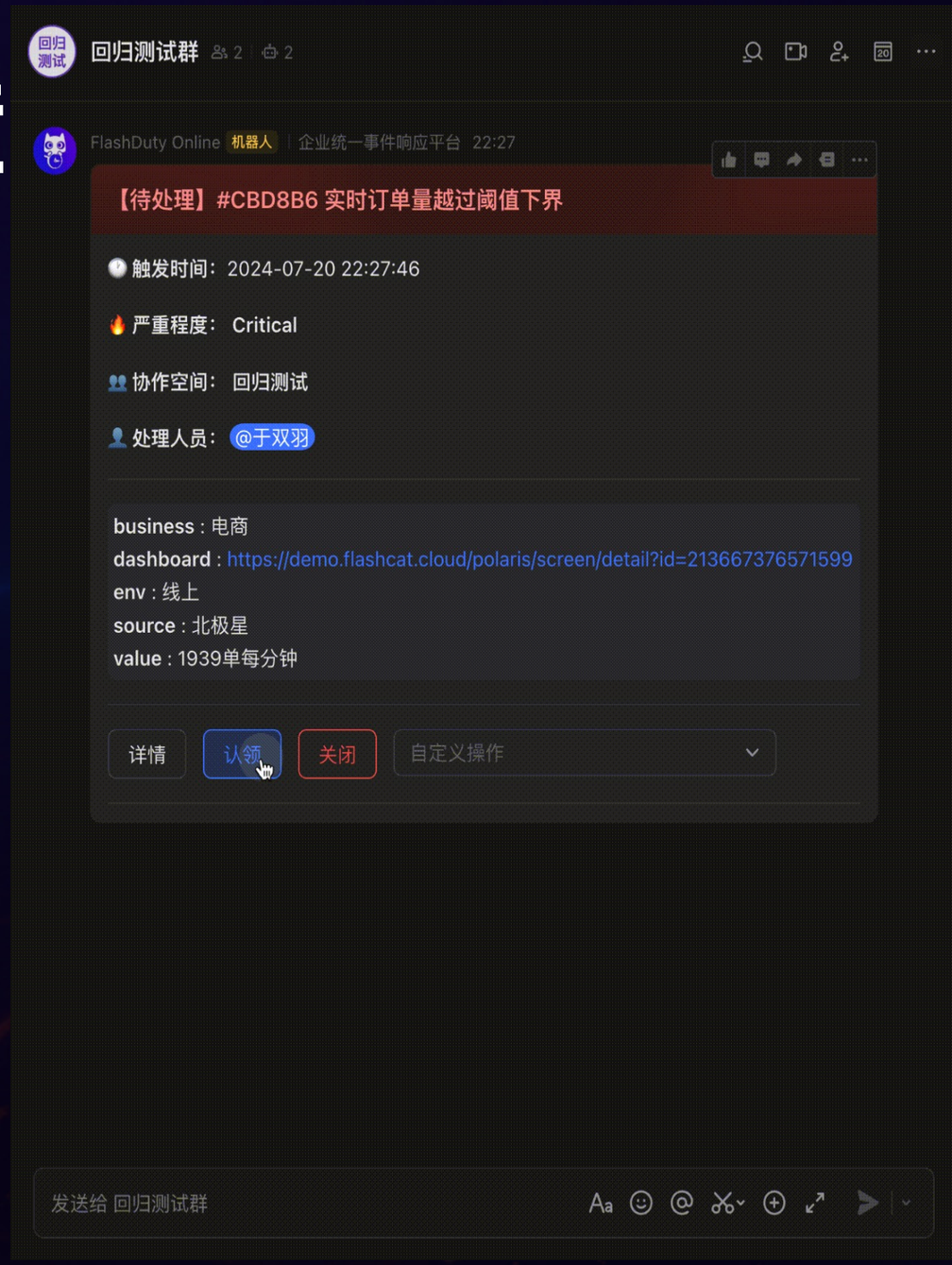
将故障的任何变化及时投递到 IM

2 **多端操作**

在 IM 内查看、处理故障，操作多端同步

3 **加强协作**

一键拉起作战室，关键信息回传到平台



中国北京 2024.7.26

自定义操作，集成工作流

API 集成

以 按钮 形式集成到控制台、IM 消息卡片

自动化流程

集成任何自动化、SOP 流程

典型场景

- 重启主机
- 回滚变更
- AI 根因分析
- 一键拉群
- 发布 Status Page

中国北京 2024.7.26

Flashduty 产品路线图

1 **故障复盘报告**

基于模板和 AI 生成故障报告

2 **智能故障分析**

智能聚合、相似故障、相关故障发现

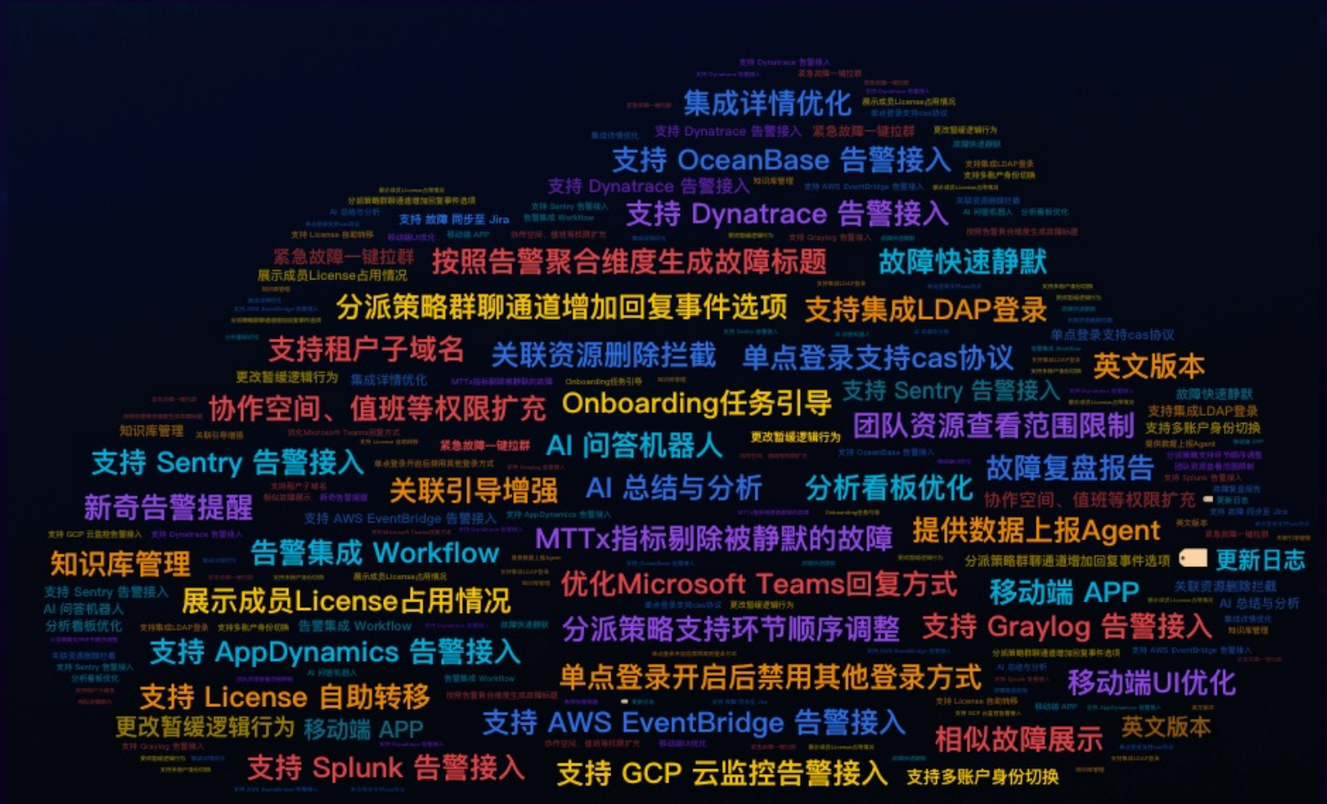
3 **移动端支持**

提供客户端app

4 **前往 Roadmap 了解更多**

http://roadmap.flashcat.cloud

中国北京 2024.7.26



感谢聆听

Thank you for **listening**

中国北京 2024.7.26