SOFAStack

高可用管理

产品简介

产品版本：AntStack Plus 1.11.0

文档版本：20221008

蚂蚁集团
ANT GROUP

法律声明

蚂蚁集团版权所有©2022，并保留一切权利。

商标声明

免责声明

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

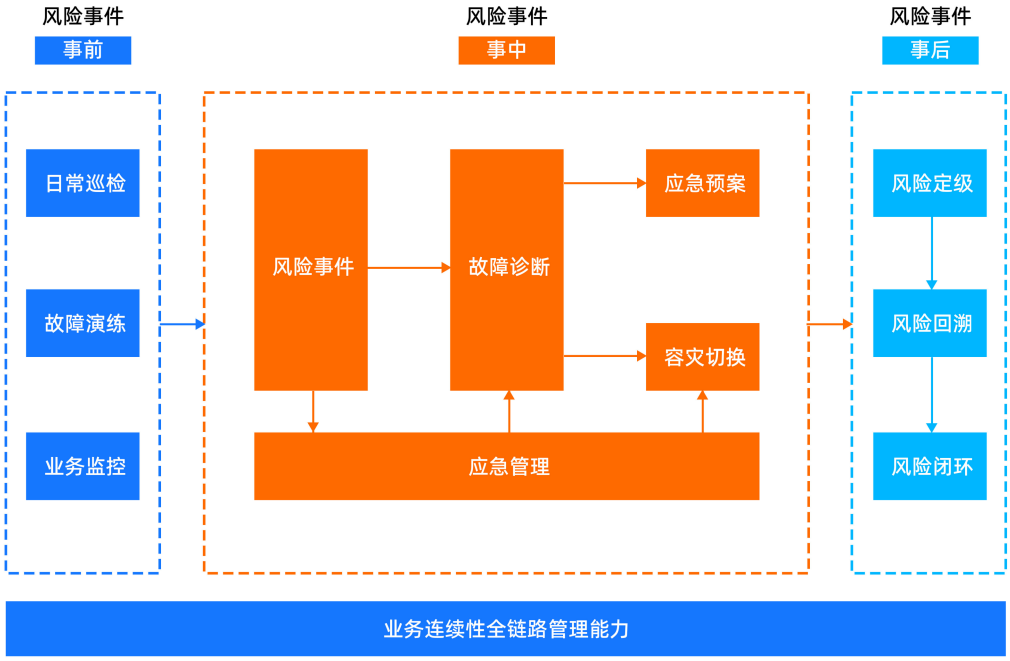
1.什么是高可用管理平台

高可用管理平台（High Availability Service，HAS）是基于蚂蚁多年技术风险防控理论和实践而衍生出的产品，是蚂蚁分布式架构 SOFAStack 的重要组成部分。

高可用管理平台覆盖了应用运行风险事件事前、事中、事后的全流程管理。

- **事前**：通过应用巡检、故障诊断以及和监控平台的联动，实现应用运行风险的主动发现。
- **事中**：通过故障诊断、应急预案、容灾切换实现风险事件快速定位和恢复。
- **事后**：通过风险定级、回溯，实现风险事件的闭环管理。有效提升 IT 技术风险防御水平，保证业务健康、持续、稳定运行。

高可用管理平台技术风险管理示意图如下：



2.产品优势

蚂蚁技术风险管理体系方法论+平台工具完整落地

通过高可用管理平台工具，可以输出蚂蚁多年积累点的技术风险防控内容，并在此基础上助力用户建设符合用户实际情况的技术风险防控体系，提升用户整体技术风险防控水平。

技术风险防控效率有效提升

通过高可用管理平台技术风险防控能力，自动化、标准化、例行化日常运维，降低操作复杂度，运维结果清晰可见，实现风险事件的闭环管理。

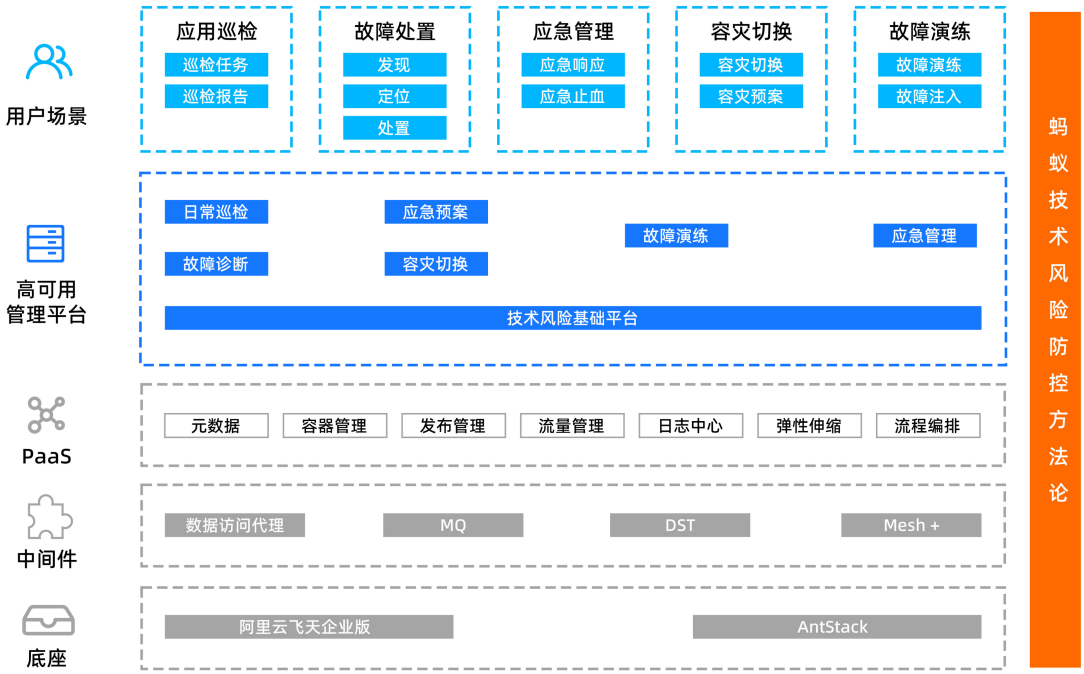
- 通过日常巡检主动感知业务运行风险，在对业务造成影响前及时处理。
- 自动化故障诊断和标准化应急预案快速定位并恢复故障，减少因故障引起的业务中断时间。
- 故障演练主动检验应用高可用能力。
- 支持蚂蚁产品双中心容灾切换，满足监管合规需求。

技术风险防控内容库快速更新

阿里云、蚂蚁技术风险团队基于域内、域外技术风险防控经验，共建日常巡检、故障诊断、应急预案内容库，用户可享受最新的技术风险防控内容。

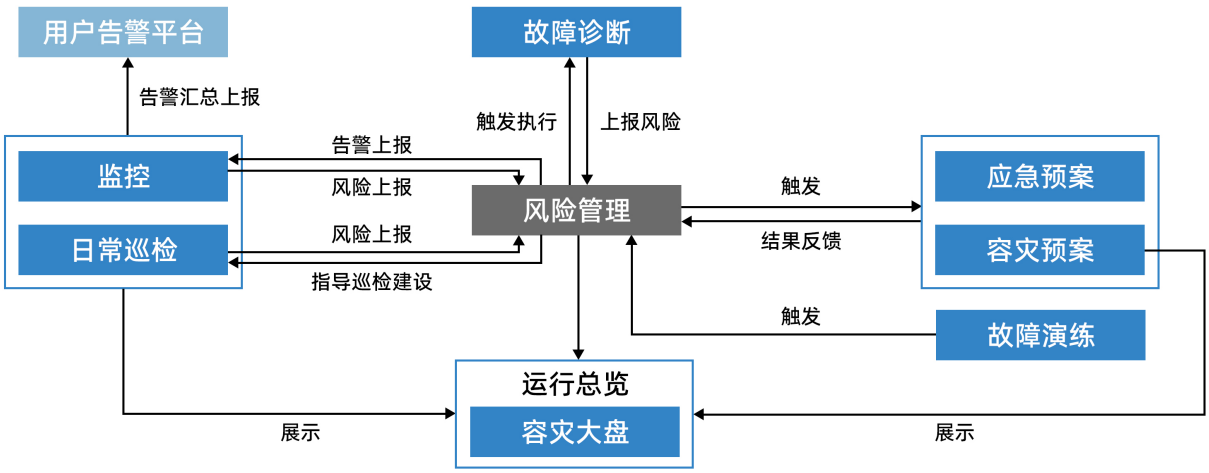
3.产品架构

高可用管理平台 HAS 在蚂蚁 SOFA 中间件以及 Café 应用发布平台的基础上，为用户应用及蚂蚁应用提供日常巡检、风险管理、应急预案、容灾切换、故障演练等技术风险管理能力，满足用户应用巡检、故障处置、应急管理、容灾切换、故障演练等多种应用运维场景。



系统架构

高可用管理平台 HAS 的系统架构如下图所示：

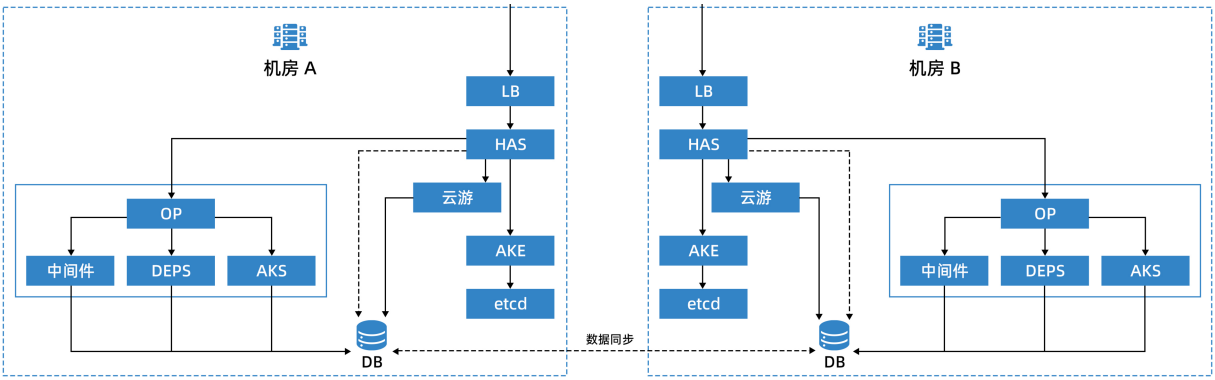


高可用管理平台 HAS 的系统架构包括：

- 风险管理模块是 HAS 的核心组件，实现风险事件的汇集以及风险事件处理的流转。
- 日常巡检、业务监控、故障诊断发现风险事件，并将其上报至风险管理中心，由风险管理中心进行统一调度，如触发故障诊断、应急预案等。
- 运行总览集中展示整个平台的运行状态、风险情况、容灾大盘。

部署架构

部署结构如下图所示。



部署依赖

HAS 部署时，需要依赖以下底层服务：

- 数据库 RDS 或 OceanBase
- 负载均衡 SLB 或 ALB
- 云游、AKE 等底座服务

运行依赖

HAS 运行时，需要借助以下平台：

- 身份与访问控制产品 IAM
- 云游

4.功能特性

高可用管理平台 HAS 定位于 IT 技术风险防控。产品主要功能包括风险管理、日常巡检、故障诊断、应急预案、故障演练等。

风险管理

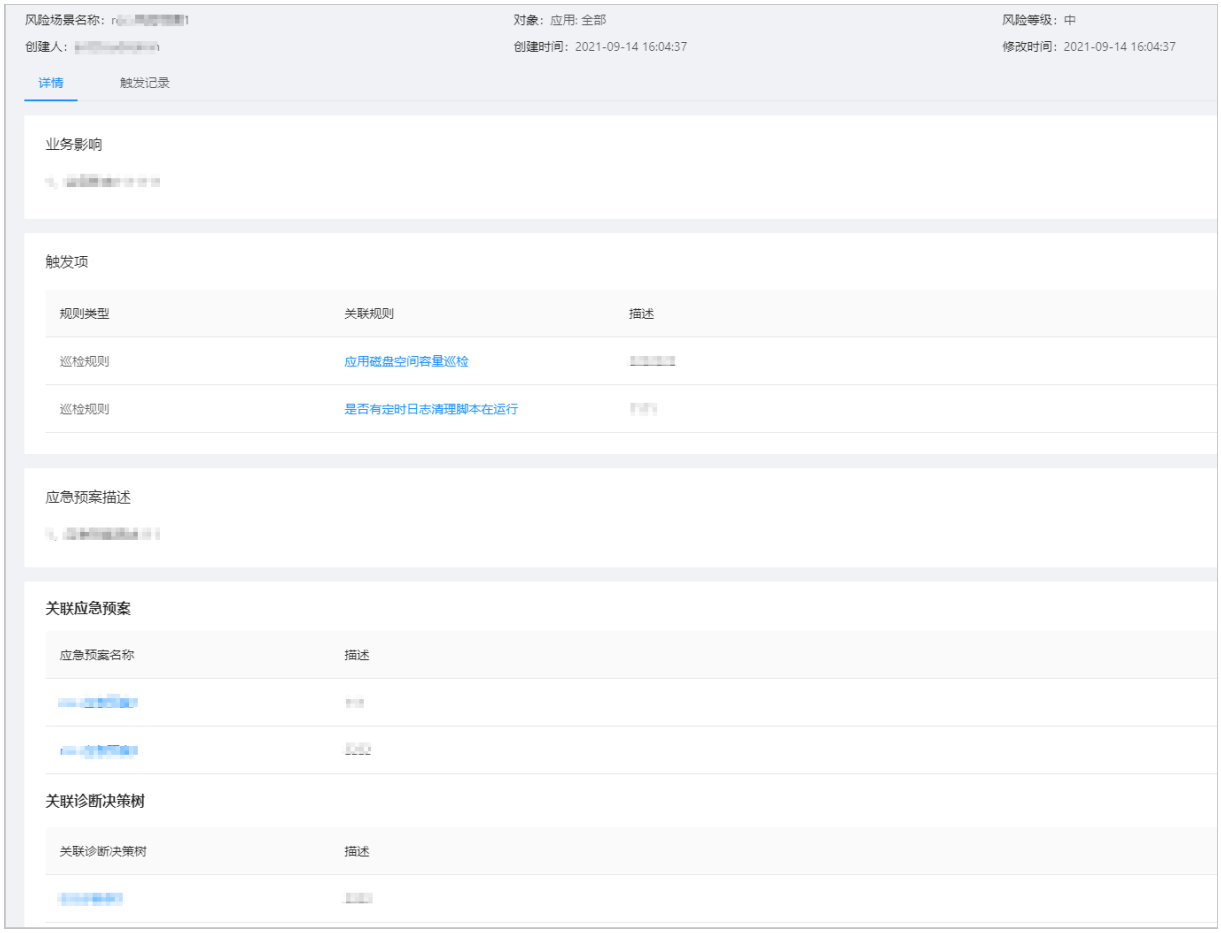
风险管理 是高可用管理平台核心，是风险事件汇集、处置的中枢平台，具体如下。

风险事件

- **风险事件汇集**：负责将监控、巡检、诊断产生的风险或告警信息进行汇总。
- **风险事件处置**：支持直接在风险事件列表中，对风险事件进行处理。在处理风险事件时，风险管理能够推荐可执行的应急预案，支持直接触发执行。

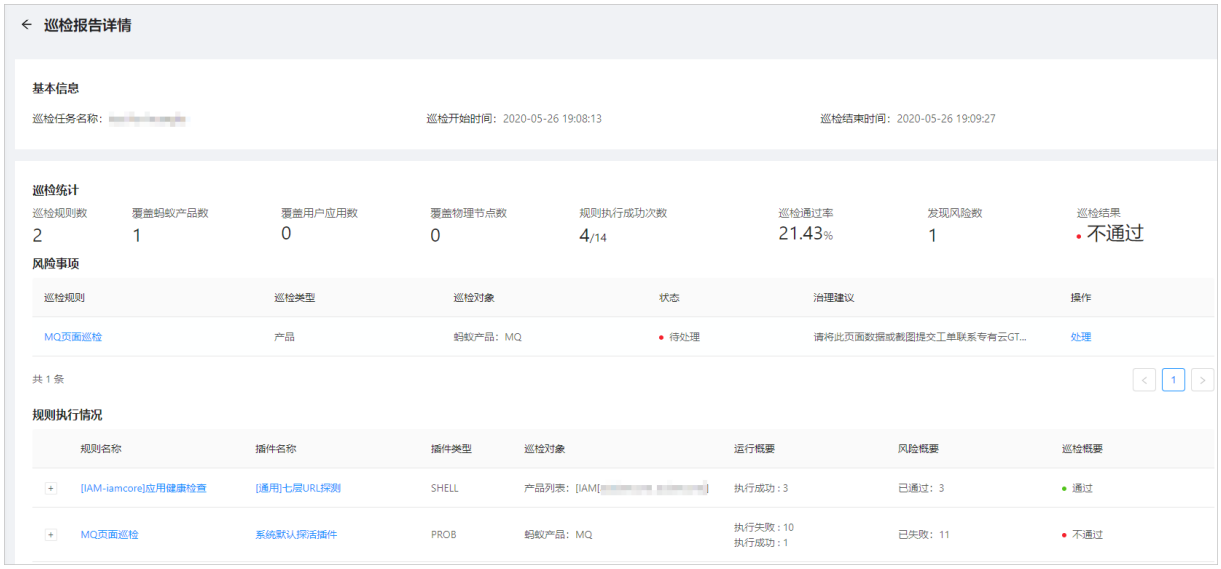
风险场景

风险场景 是针对特定风险事件进行集中化处理的模块，风险场景中包含了处理风险事件所需要诊断决策树、应急预案、业务影响等信息。目前应急场景升级后，需要将风险场景和应急响应联动，所以需要添加更多属性。



日常巡检

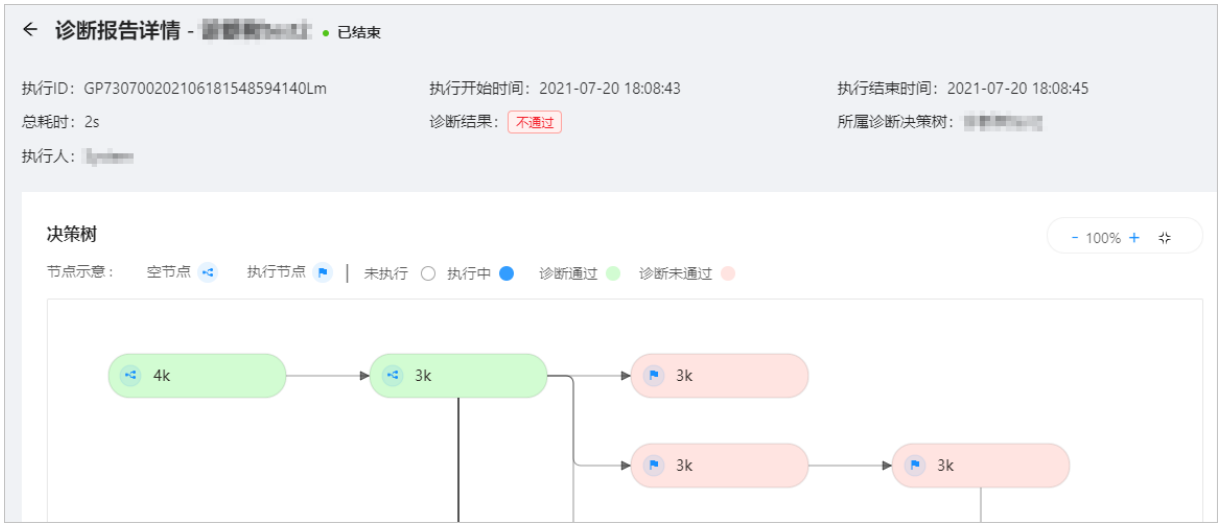
日常巡检 是高可用管理平台最常用的功能。通过日常巡检功能，可以例行化、自动化地对系统稳定性、可用性进行巡查，并将巡检结果实时同步推送至指定的钉钉群中，便于运维人员第一时间了解应用风险；同时支持生成巡检报告，供运维人员统一归档。巡检插件支持多种类型，包括 Java、python、shell、自动化测试镜像、页面探活等。用户可以根据应用情况自定义巡检插件。同时，高可用管理平台也提供了蚂蚁内部及各个用户长期使用过程中沉淀的巡检规则，开箱即用。



故障诊断

故障诊断 的核心能力是将在运维人员头脑中或在故障排查手册中的经验、排查过程通过工具平台进行沉淀和展示。

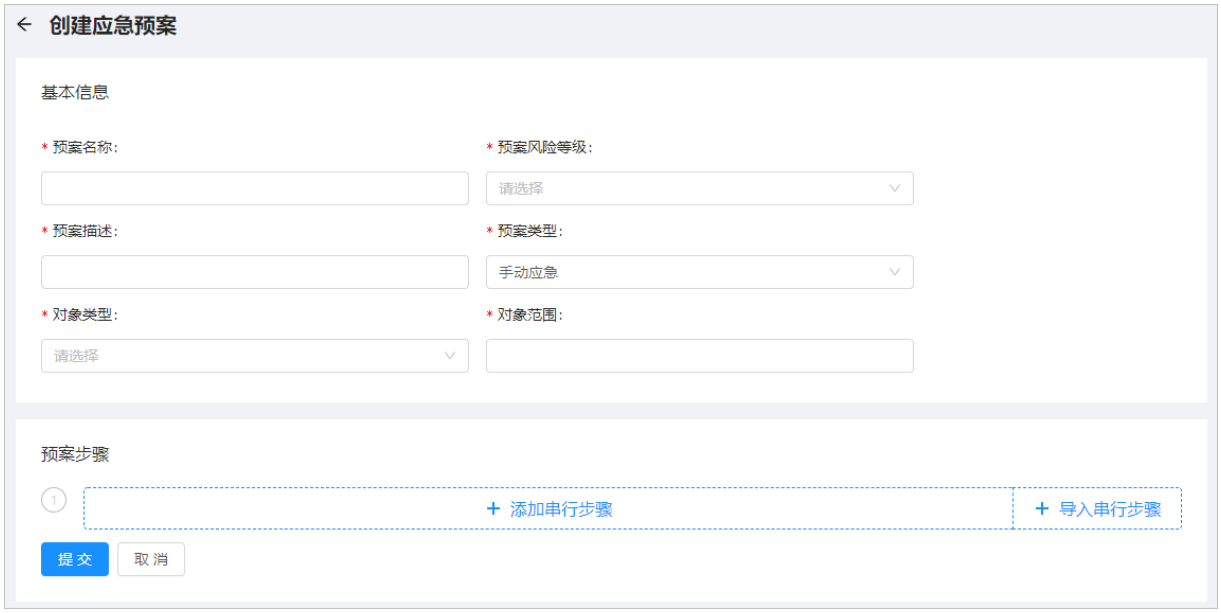
运维人员通过决策树方式，图形化地编排故障诊断过程、设计排查顺序。继而在风险事件发生时，将例行化、程式化、标准化的排查过程，通过故障决策树自动执行，并直接反馈诊断结果。通过故障诊断平台，能够极大地缩短故障排查时间。同时，屏蔽了不同运维人员在故障排查时的经验和技能差异，实现故障的快速定位。



应急预案

应急预案 提供了应用运维原子操作的编排能力，如应用重启、应用摘流、数据库切换、物理服务器重启等操作。

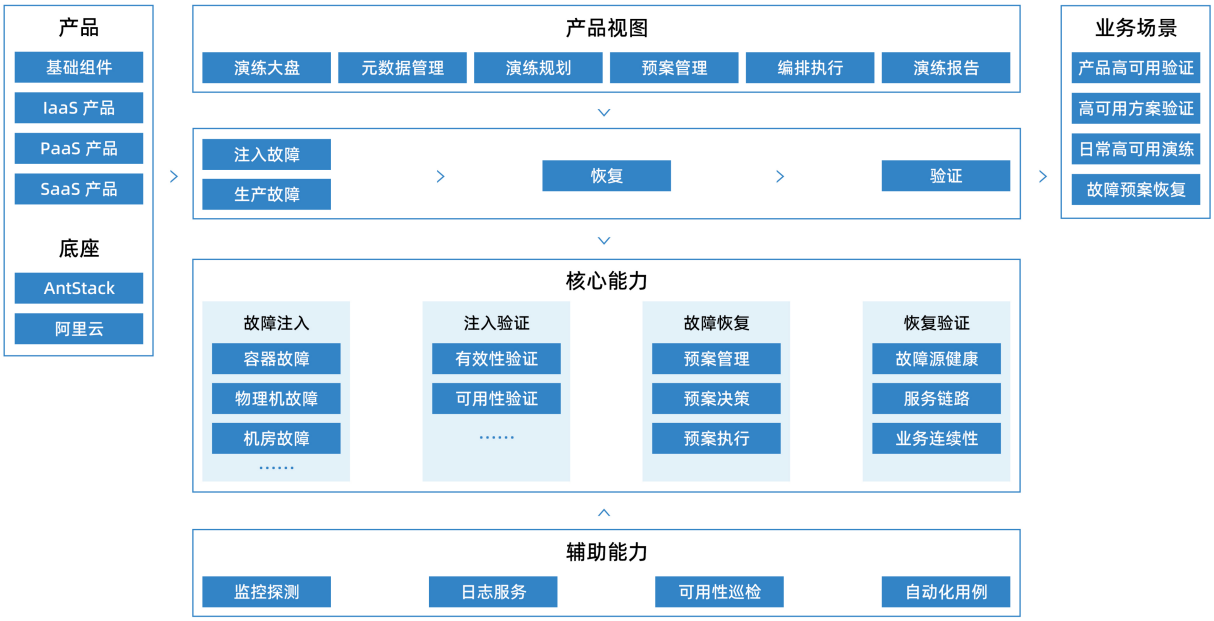
运维人员可以根据常见故障场景的处理过程，选择需要的原子能力进行编排组合，形成可执行的应急预案。当风险事件发生时，风险事件中心会推荐可执行的应急预案，供运维快速选择并自动化执行，从而通过标准化处理流程，实现故障快速恢复。



故障演练

故障演练 提供了故障注入能力，通过演练平台主动触发故障，以此观测应用软件的高可用性。

故障演练平台支持触发 CPU 利用率升高、内存利用率升高、内存占用、网络丢包、容器宕机、物理机宕机等常见故障，并针对故障制定出详细的演练和恢复计划，保证用户能够有计划地测量和观测应用高可用能力。

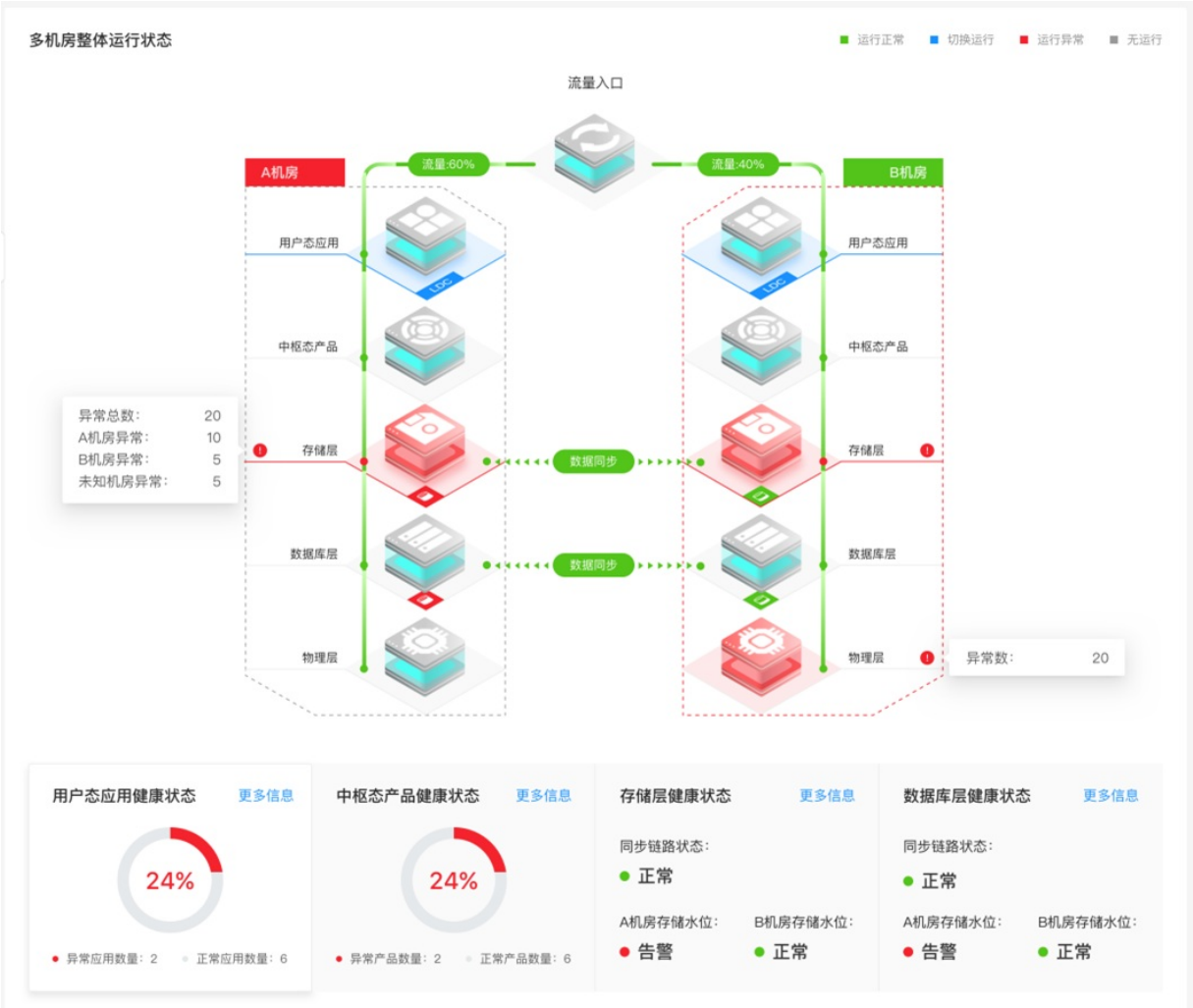


容灾管理

容灾大盘

容灾大盘 为用户提供了一个全局视角，可查看整个多机房容灾架构图，以及多机房的运行状态情况，方便用户判断容灾的整体运行情况、是否具备可进行容灾切换的条件以便进行问题排查。

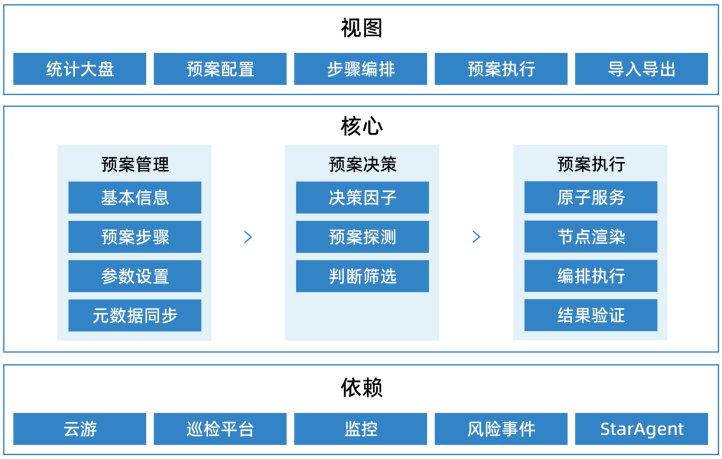
容灾大盘架构主要分为业务层、中枢层、存储层、数据库层、物理层等，分别展示不同层面的应用异常和告警情况。其中存储层和数据库层，除了应用异常监控外，还包括数据同步链路以及容量水位情况的监控。



容灾预案

容灾预案 是基于产品、站点运行期以及容灾切换中沉淀的相关故障应急措施，提供一站式的预案管理、预案决策以及预案编排执行的产品能力。具备以下几个特点：

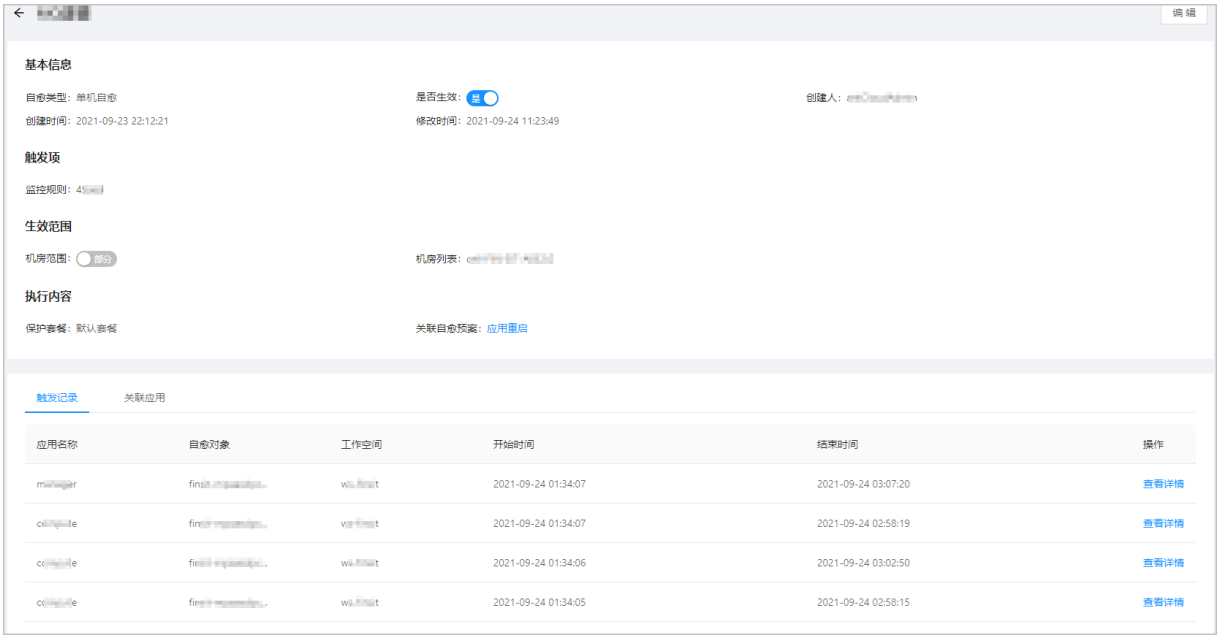
- **可复用**：历史问题、通用问题的解决方法和应急方案可以不断沉淀积累，并在各个站点复用。
- **透明化**：应急处理过程、处理方式、做了什么变更等一目了然。
- **高效率**：提高应急处理效率，尽可能减少故障恢复的时间。原来需要根据文档判断并按照文档手动变更的方式，变为根据场景定位到预案后直接一键执行。
- **降成本**：把应急处理措施预案化，降低学习门槛、人员投入成本、站点维护成本等。



风险自愈

为了提高应急效率，**风险自愈** 为用户提供高级的故障自愈能力，将特定场景下的处理过程进行固化，在故障发生时，自动执行应急预案，完成故障的自愈处置。

风险自愈平台将指定场景下的处理过程进行编排，并和监控、巡检功能对接，当在正式环境中，巡检或监控告警触发后，风险自愈平台按照编排好的预案，在保护规则的作用下自动执行，达到风险故障自动化处置的目的。



5.应用场景

本文主要介绍高可用管理平台 HAS 的应用场景。

日常风险防控

在日常的运维场景中，通过多功能模块的联动使用，自动化执行日常运维脚本，实现定期可控的日常巡检运维；同时不断更新优化日常巡检、故障诊断、应急预案等内容的建设，不断丰富和完善应用技术风险防控体系，简化日常应用运维操作。

故障演练

为不断提升产品高可用能力，通过高可用管理平台的故障演练模块，设计并规划演练计划和恢复方案，继而在演练过程中不断发现、解决容灾预案存在的问题。以此，降低产品使用过程中故障发生概率，提高故障恢复效率，进而实现产品高可用性的有效提升。

机房级容灾

- **同城双活**

  同一个城市，建设两个机房环境，两地距离 50 km 以内，万兆光纤专线互连，业务应用层面可以两个机房同时提供业务服务，当一个机房故障，不影响另外一个机房业务使用。

- **异地主备**

  满足容灾需求，两地不同城市分别建设两个机房，一主一备，两地距离超过 1000 km，主机房承载业务流量，备机房无业务流量，只做备用机房使用。当主机房故障，可以切换流量到备机房快速恢复业务，等主机房故障恢复以后，再回切流量到主机房。

- **两地三中心**

  两地三中心，也称为同城双活加异地主备方案，即上文提及的同城两机房做双活部署，外加一个异地机房只做备份，不承载任何业务流量，基本只做冷备数据使用。最大程度地保障了数据的高可用备份。

- **LDC 单元化（异地多活）**

  LDC 单元化架构是可以实现异地多活和高并发场景的架构体系，LDC（Logic Data Center）逻辑数据中心是相对于传统的 IDC（Internet Data Center）提出的。逻辑数据中心所表达的中心思想是无论物理结构如何的分布，整个数据中心在逻辑上是协同和统一的。主要适用于大型互联网公司在线交易系统支持，比如淘宝、支付宝、携程等。

6.基础术语

AKE

容器引擎（Ant Financial Kubernetes Engine，AKE）是将底层物理资源按照计算、网络、存储等进行切分和抽象的容器引擎。AKE 通过使用 Kubernetes 和 Docker 技术将整个物理资源进行池化，向上层服务提供按量使用的计算、网络和存储资源。

ALB

负载均衡（Ant Financial Load Balancer，ALB）是将访问流量根据转发规则分发到后端多台后端服务器的流量分发控制服务。通过流量分发扩展应用系统对外的服务能力，通过消除单点故障提升应用系统的可用性。

IAM

蚂蚁科技身份访问管理（Identity and Access Management，IAM）控制台是管理成员、分配权限、管理身份源、查看操作记录的平台。

OceanBase

OceanBase 是阿里巴巴与蚂蚁科技独立自主研发的一款分布式关系数据库产品，融合传统关系数据库和分布式系统的优势，具备高可用、高性能、高可扩展性，在功能上兼容 MySQL 等特点，在通用硬件上提供金融级高可用的数据库服务。

RPO

数据恢复点目标（Recovery Point Objective，RPO），以时间为单位，即在灾难发生时，系统和数据必须恢复的时间点要求。RPO 标志系统能够容忍的最大数据丢失量。系统容忍丢失的数据量越小，RPO 的值越小。

RTO

恢复时间目标（Recovery Time Objective，RTO），以时间为单位，即在灾难发生后，信息系统或业务功能从停止到必须恢复的时间要求。RTO 标志系统能够容忍的服务停止的最长时间。系统服务的紧迫性要求越高，RTO 的值越小。

容灾预案

指包含容灾步骤的可执行预案。

页面探活

指通过浏览器打开巡检页面来判断页面存活情况。高可用容灾平台除了支持无需登录的静态页面探活外，还支持需要登录态的页面探活，也支持匹配页面的内容或元素来确定页面是否已渲染成功。在高可用容灾平台上，可以将页面探活配置成巡检任务以定时巡检页面。

云游

云游是蚂蚁科技的一站式专有云规划、交付、运维平台，管理着专有云从诞生到落地的整个生命周期。