智能运维系列（六）| 如何通过智能化手段将运维管理要求落地？

智能运维系列专题简介：智能运维（AIOps），根据 Gartner 的最新阐释，意指整合大数据和机器学习能力，通过松耦合、可扩展方式去提取和分析数据量（volume）、种类（variety）和速度（velocity）这三个维度不断增长的 IT 数据，进而为 IT 运维管理产品提供支撑。在此，微众银行智能运维团队根据一线工作的实践经验与心得体会，特别撰写了《智能运维系列》文章，敬请持续关注。

点击回顾：[智能运维系列（一） | AIOps 的崛起与实践](https://www.infoq.cn/article/fqUfkjhecOla1zKUKycN)

点击回顾：[智能运维系列（二）| 智能化监控领域探索](https://www.infoq.cn/article/Qta6VCyjvHdoiJg5wKze)

点击回顾：[智能运维系列（三）| 浅析智能异常检测——“慧识图”核心算法](https://www.infoq.cn/article/mryjNLXOlujV7fkQFUaL)

点击回顾：[智能运维系列（四）| 曝光交易路径](https://www.infoq.cn/article/a72WeEMuM1O63iX1w0ZK)

点击回顾：[智能运维系列（五）| 浅析基于知识图谱的根因分析系统](https://www.infoq.cn/article/cUYWKqYxrBamV7GwVVjM)

保障业务的稳定运营一直是IT 运维人员的首要任务。为了实现这一目标，微众银行正式启动了”异常事件根因分析 (RCA) 项目”。经过包括运维人员在内的多个角色的共同努力，最终该项目取得较为理想的效果。作为一线运维人员，笔者从运维管理的目标切入，介绍如何通过智能化手段将运维管理要求落地的实践过程，期望能给读者们带来新的思路与深层思考。

**异常检测：快速、准确识别异常**

**管理目标**

IT 异常事件管理的首要目标是快速发现业务异常。微众银行异常事件管理规范中对异常事件从影响的对象、用户量、时长等方面给出明确定义和分级，并要求在 10 分钟完成异常事件的通报。

具体到项目目标中，就是对关键产品的异常能够做到快速发现，首先有异常要能快速识别到并通报出来，同时避免过多误报对运维人员造成骚扰。此外，需要使用智能化识别方法和手段，摒弃规则、阈值等各种繁杂的人工配置方式。

总结起来就是两个字：“快”、“准”。

**具体实践**

虽然以上管理要求简单明确，然而现实情况却相对复杂。由于应用监控、基础组件监控分别由不同的运维团队独立维护，很多情况下需要依赖人工识别和收集，因此常规操作无法满足 10 分钟完成通报的要求。

对于异常检测来说，我们要求自动发现业务产品交易行为的异常表现。首先确定异常检测范围，关注业务产品的关键场景，特别是可能影响产品可用性的交易，如贷款产品的登录、开户、借款、还款等。其次，需要通过机器学习算法识别这些交易场景的成功率、交易量、平均时延三种指标在一段时间内的波动是异常的，自动推送消息到微信公众号或微信群，实现异常的自动通报。

要做到及时、准确，不漏报、也不误报绝非易事。异常检测和使用渔网捕鱼类似：渔网网眼太大，会有漏网之鱼，导致有些异常识别不到；渔网网眼太小，可能会网起一堆泥沙，效果和一般的同环比监控策略没有差别，让真正需要关注的异常淹没在大量异常预警中。

针对上述技术难题，我们采取了以下措施：

1、对纳入检测范围的交易场景，我们根据发生频率即单位时间内的交易量大小，用聚类算法分为高频、中频、低频三类交易，并针对不同频率的场景指标，采用了不同的识别灵敏度和不同的通知策略。如果高频出现异常立即通报异常；对于低频异常，可以持续检测多个异常点后再通报。

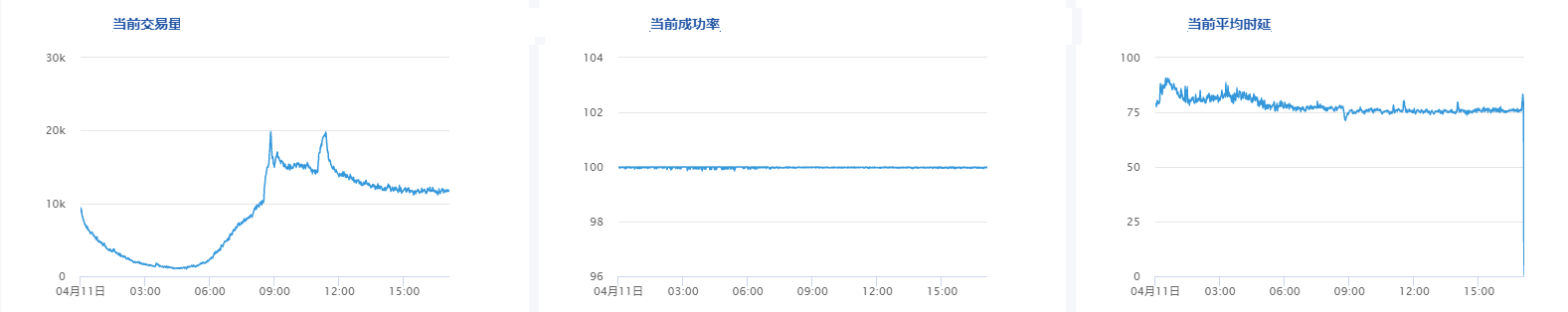


图 1 高频交易场景指标

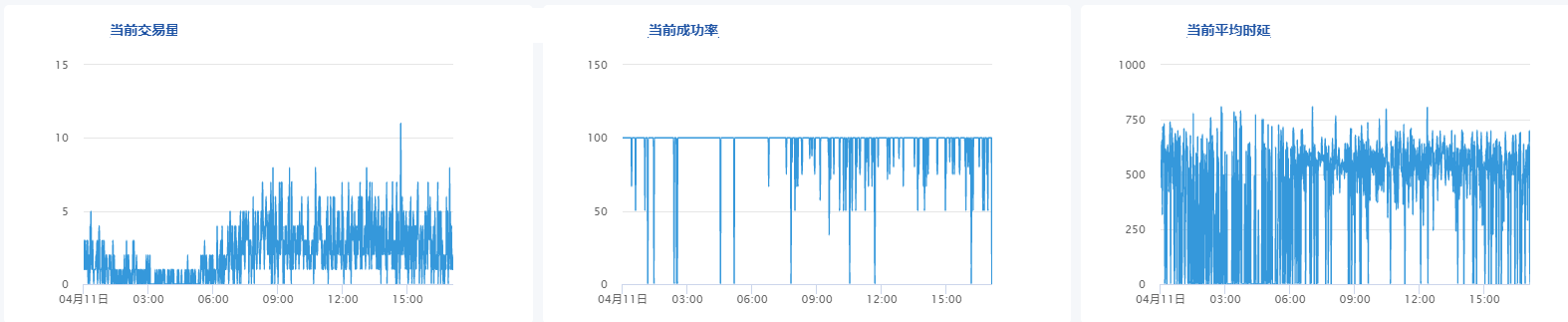


图 2 低频交易场景指标

2、根据业务特性，不同时段的交易频率采取不同的针对措施。有些产品在工作日交易量较大，非工作日交易量较小。因此，我们将多个指标的曲线做关联分析，如在分析成功率异常时，同时分析交易量指标，只有在失败的交易量达到一定的级别后才升级为异常。

3、对检测到的异常进行分级。根据产品或场景重要程序确定各个异常的关注度程度，以触发后续的通知机制和响应机制。

4、对不同指标曲线的特征采取不同的针对措施。规律性强、取值范围较稳定的指标，异常识别会更容易和准确一些；而波动较大的曲线异常识别难度则相对较大。在实践中，我们一方面需要分析指标波动是否合理，另一方面还需要分析异常识别结果的有效性，对异常识别算法需要不断调整和优化，从而减少误告或漏告。

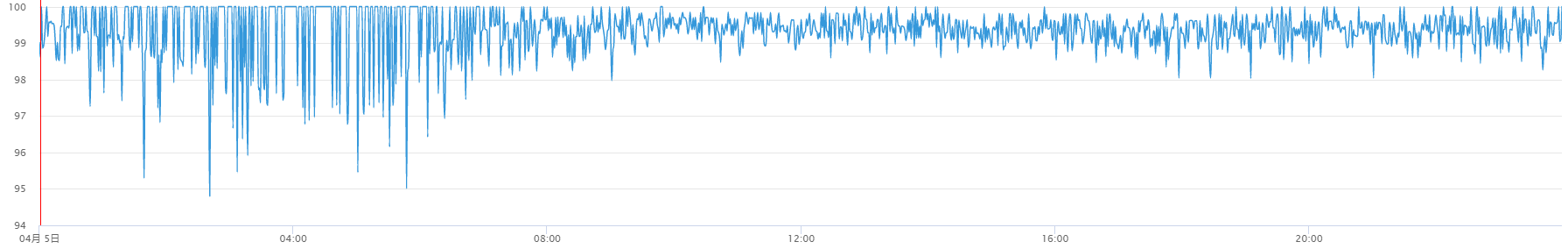


图 3 较稳定的曲线（单日趋势）

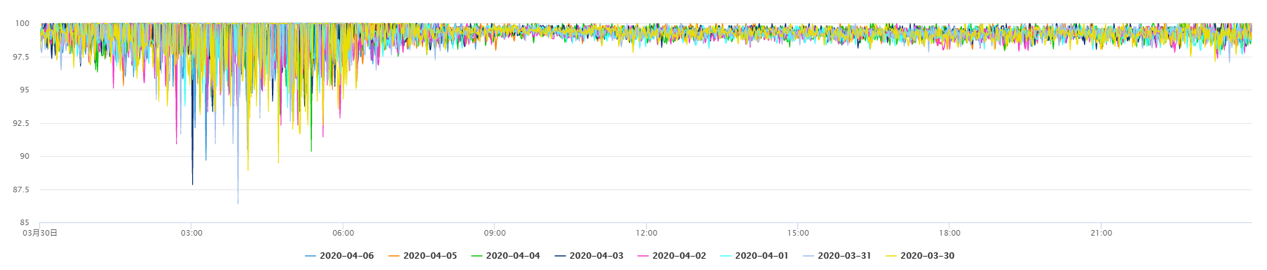


图 4 较稳定的曲线（多日趋势对比）

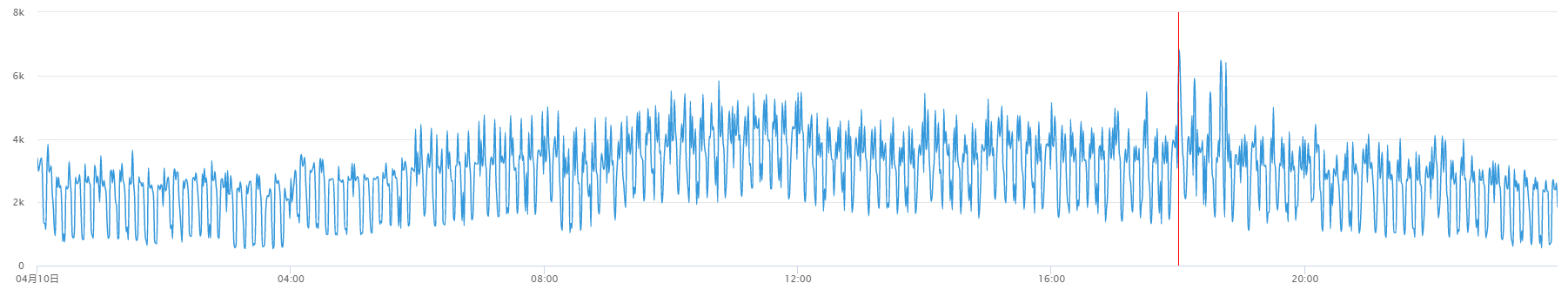


图 5 波动较大的曲线（单日趋势）



图 6 波动较大的曲线（多日对比趋势）

5、其他注意事项。

* 算法并不能解决所有问题，对于特殊曲线，应根据实际的运维需要将自动识别算法和规则相结合。
* 动态快速学习，人为标记后能够快速动态调整，如标记为误告后类似波动不再识别为异常。
* 异常检测应支持灰度，在累积一定的历史数据，或观察异常检测准确率达到期望范围内再开放给用户。

**实际效果**

通过采用以上措施，我们已做到异常的自动识别和自动通知时间控制在 3-4 分钟内。另外，针对当前的一些高频交易曲线，我们已经实现了秒级检测，异常的自动识别和自动通知时间从 3-4 分钟缩短到 30s 以内，微众银行异常事件主动发现率已达 95% 以上，平均通报时长也较之前缩短 50% 以上。

**异常事件根因定位：快速、准确判断异常原因**

**管理目标**

出现异常后，需要快速响应并处理异常，以降低对 IT 服务影响。因此，异常事件 30 分钟内恢复率也是我们持续监测的管理指标之一。

要快速恢复，多数情况下需要先找到引起异常的原因。以往运维人员要通过人工排查告警、回顾变更记录、查看日志等多个方面来寻找原因，一般需要多个组件的运维和业务运维共同参与，信息的沟通和排查需要较长时间。从我们对异常事件处理过程的持续分析来看，异常定位环节耗时占比最高。

针对以上问题，我们的解决之道是通过智能化手段快速、准确找到引发事件的真正原因，为后续事件恢复争取宝贵的时间。

**具体实践**

检测到异常后，我们需要快速收集业务服务调用链上各节点的异常证据，进行根因分析。另外，我们还需要匹配是否曾经发生过类似事件，给出历史根因信息供参考。

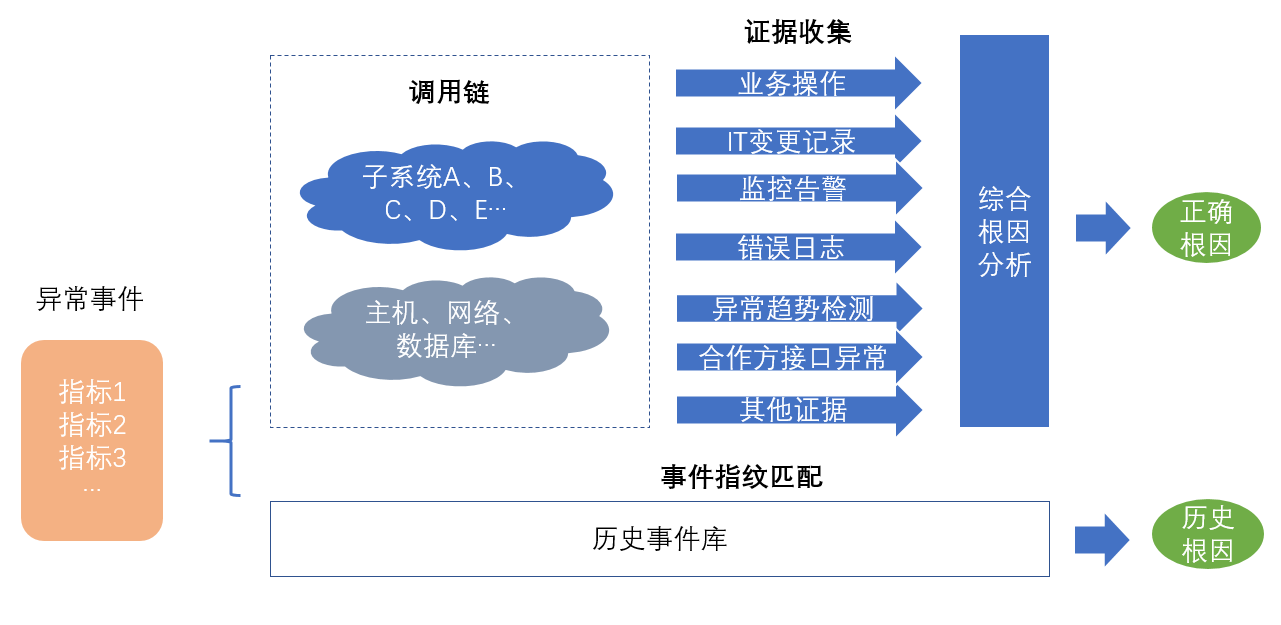


图 7 根因分析过程示意图

调用可以分为两个维度：横向调用、纵向调用。横向调用是指对从入口子系统开始至后端所有交易路径上经过的所有子系统进行分析；纵向调用指对支撑应用层子系统运行的所有逻辑层、物理层，包括数据库、网络、主机等进行分析。

微众银行大部分子系统之间的相互调用使用统一的消息总线，因此可以通过消息日志自动分析横向调用关系。未通过消息总线的调用也基本上有 CMDB 记录，可通过 CMDB 中记录的各类 CI 之间的关系进行分析。

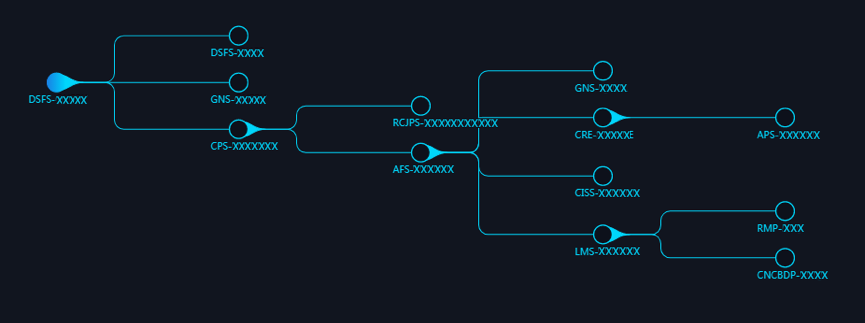


图 8 某个业务场景的横向调用关系图

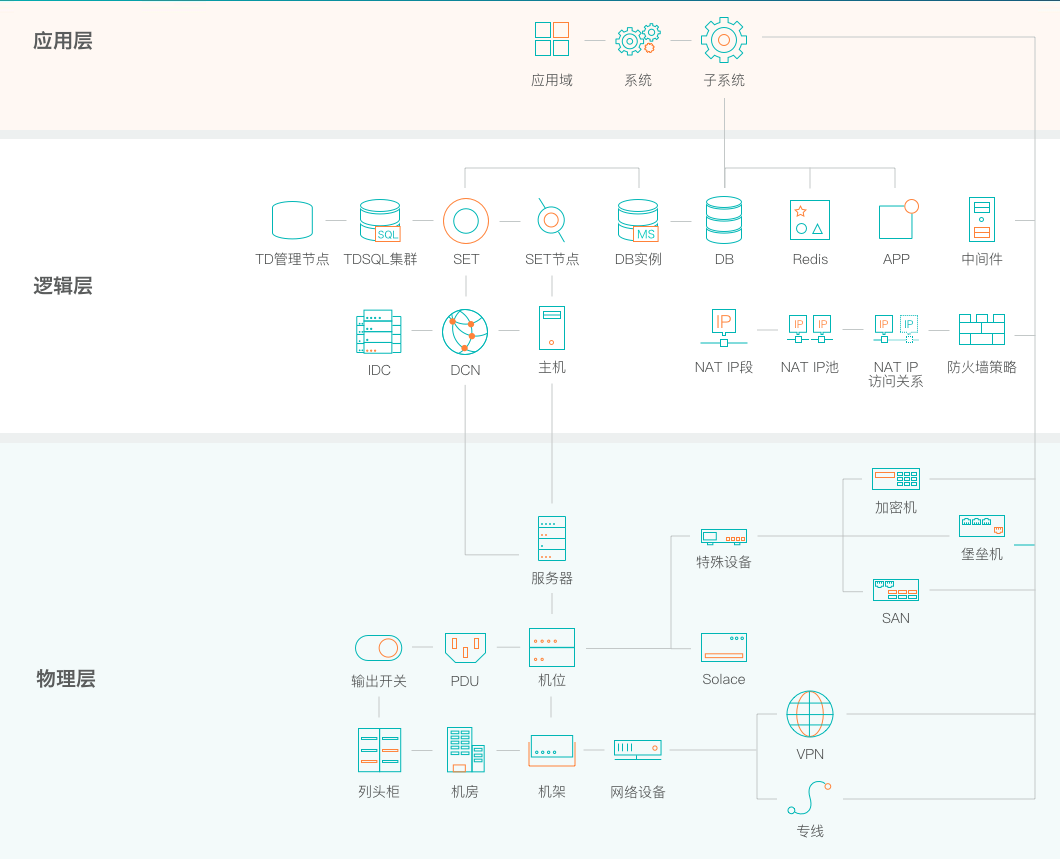


图 9 纵向调用示意图

证据收集过程中遇到的最大挑战是保证证据的全面性，避免过于依赖单方面信息造成综合根因分析的误判。应根据调用链的每类节点分析可能存在的异常表现和原因，并对这些信息进行全面收集和相互佐证。如数据库异常的表现包括应用日志有错误信息、在监控系统中有告警、数据库的 CPU/IO 等曲线有异常波动等。若仅仅依赖数据库告警来判断 DB 是否有异常，当告警配置有缺失时可能就无法识别到正确原因，此时还应该主动对数据库指标曲线做异常趋势监测。

变更及 IT 操作也是造成异常的主要原因之一。应收集每类可能带来业务影响的变更操作的记录，包括停服维护、版本发布、底层基础架构变更、业务人员 / 运维人员批量操作、外部变更等；收集方式可以采取变更过程中自动上报、主动采集等方式。

**实际效果**

在异常发生后 2-5 分钟内自动推送根因分析结论，该功能给运维人员带来很大的便利，在确认原因属实后，可以快速实施相应的恢复方案，异常恢复时效自然得到很大的提升。经过近一年的不断优化，微众银行根因定位准确率趋势提升至 80% 以上，将近 70% 的异常也都能在 30 分钟内恢复。

**建立持续优化机制**

为了衡量异常事件根因定位的实际效果，我们重点关注两个指标，异常识别准确率、根因分析准确率，并作为我们持续优化的目标。每次异常发生后，运维人员需要确认实际根因，并且通过复盘，从以下维度对异常进行检视和分析：

* 是否及时识别到？未识别到的原因是什么？需要如何调整保障下次能识别到？
* 根因定位是否准确？为什么没有定位到正确的根因？缺失的证据如何才能收集到？
* 业务层面应该做哪些优化？

通过以上检视和分析，我们会形成需要进一步跟进和落实的行动项，可能包括算法优化、业务逻辑优化等多个方面，进而持续跟踪直至解决。正是通过这种良性的持续优化机制，我们才能推动根因定位准确率和效率不断提升。

综上所述，通过智能化手段的运用并持续优化，微众银行实现了异常事件快速发现和恢复的目标，极大提升了运维工作的效率。如需了解我们在异常检测和根因定位中使用的机器学习算法，请参阅该系列其他文章。