智能运维系列（二）| 智能化监控领域探索

**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" \t "_blank)

**缘起**

最近几年，随着大数据的沉淀以及AI 技术的兴起，智能化运维AIOps 的概念逐渐成为一股热潮。然而，真正把智能化做起来并起到实际的效果，对每家公司来说，都是不小的挑战。回想微众银行启动智能化的初衷，并非单纯追逐“AIOps”这个热潮，而是遇到了要解决的问题，且基础数据积累已初具规模，算法基本能满足要求，才真正行动起来去做一些尝试。这个尝试也不是一帆风顺，其中经历了种种困难和挫折，经过2 年多的努力截止到2020 年初终于有一些实质性的收获。

2017 年年初，经过 3 年多的努力，微众银行各类基础运维工具第二代都顺利搭建完成。CMDB、ITSM、自动化发布工具、监控系统（简称 IMS）、容量系统（简称 ICP）都可以正常运作，特别是 CMDB 的推广使用效果非常好。然而，运维工具同时面临两个压力，**一是生产异常如何快速识别和定位，二是运维工具团队的产品未来该如何发展和进步，从而保持行业内领先。**

虽然微众银行的监控系统监控覆盖面足够，但监控信息都是以点的状态呈现。运维团队曾经尝试通过 CMDB 的关系做告警压缩，然而效果并不明显，运维活动仍面临以下几个具体痛点：

* **大量告警，无法快速获取影响范围，异常升级速度很慢：**负责人需要找各个产品运维的同事确定影响范围（微众银行以产品为视角进行监控，监控时会细化到各个子系统），曾出现过负责人把各个产品的运维负责人召集过来挨个询问的情况。
* **缺少系统化的处理流程和方法：**异常出现时，运维同学常常手忙脚乱的处理具体问题，缺少一套系统化的流程和方法来实时地监控异常并排查可能存在的根因。

正是由于以上两个问题，在进行异常复盘时，运维团队常常受到 CIO 质疑和挑战，理由是监控系统没有展示清楚。

微众银行分布式架构在行业内知名度很高，近几年得过很多奖项，与此同时，运维管理系统需要进一步升级来与之匹配。一个好的运维系统不仅需要满足日常工作，还需要持续提升运维效率，降低运维人力投入，持续不断地将运维人力投入成本降到最低。**CIO 曾在 2015 年对运维工作提出了相关要求：“监控系统能否和阿尔法狗一样，自我学习和进步，定位根因？”**虽然当时运维工具还在处于较为初级的阶段，然而这个要求却为未来微众银行智能化运维的发展指明了方向。

针对上述问题，运维团队最初涉及了一个相对简陋的解决方案：建立一个简单的规则进行对比，不管现在是正常还是异常，默认计算当前生命曲线的值的异动情况，系统默认设置了一个基数值，直接判断与该基数值的差异，来进行预警。诚然，这个设计距离真正的智能化运维还有较大的差距，充其量只能说是做了个简单的报表以解燃眉之急。因此，微众银行在过去几年进行了持续的积累和优化，下文将为大家具体说明。

**运维数据的积累：标准化和规范化**

15 年至 17 年微众银行在运维标准化、自动化上取得较好的进展，为运维数据的积累打下坚实的基础：

* CMDB 全面推广并运作良好：有机制保障 CMDB 信息准确性，运维同事信任并使用 CMDB 信息，CMDB 系统自身提供大量 API 供外围系统使用 ;
* 监控数据的积累：从基础架构层到应用层监控数据记录在监控系统 IMS 中 ;
* 日志错误类：将日志中错误类信息统一采集和存储（错误日志有关键字告警）;
* 自动化发布平台：记录产品运维人员应用发布，数据库操作等各类实时操作记录 ;
* 基础架构 WeCloud 平台：记录基础架构层（如网络，数据库底层，中间件，主机等等）实时变更、操作记录 ;
* 业务推广信息获取。

**初探：根据场景研究算法的落地**

真正的智能化探索工作，始于 2018 年年初，先从以下两个方面开始实际工程的开发工作，这两项工作对后续的异常定位和根因分析起到了基石的作用，如果这两项工作没有开展，异常快速识别和根因定位基本上无从谈起。

**（1）对微众银行“关键产品黄金生命指标曲线”进行自动检测，无需人工设置阀值，系统自己计算阀值区间，自动告警（该功能称之为“识图”，详细的介绍文档请参见后续文章《慧识图》）**

该功能的上线，标志着微众银行开始将管理监控（从全行重要产品的功能出发）的重点逐步从细粒度的底层告警收敛走向智能化的识别异常产品黄金生命指标的健康状况，从而快速获取异常指标的范围。当然这并不会改变监控系统日常告警的处理，一线人员还是会按要求进行告警的实时跟踪，监控系统同时也会提供 API 给各产品部业务运维人员进行告警的二次收敛和分析。

针对产品黄金生命指标的监控，传统方式是使用同环比监控，需要运维人员根据经验设置阀值，各产品部门的做法不一，识图上线后从监控配置工作量的下降到告警准确度上都有一个质的提升。

**（2）将服务治理的数据旁路出一份，根据业务流水号，时间点将各条零散的数据组成交易树，对交易树进行实时检测并针对中断情况进行告警（该功能称之为“Knowing 诺音”，详细的介绍文档参见后续文章《曝光交易路径》）**

产品的关键功能的每一笔交易，都可以通过流水号生成唯一的交易树，交易树在日常监控和运维管理中起到非常重要的作用。”Knowing 诺音”通过 LSTM/ 深度神经网络对每一笔流水进行实时检测，实时发现生产中的交易异常情况并生成告警。

在进行实时检测算法上，微众银行与清华大学裴丹教授进行了深度合作，裴丹教授是我国在智能化运维领域研究方向的领军人物。该合作使得微众银行在智能化运维的方法论和实践上都有了长足的进步。

当然服务治理的 trace 链并不能涵盖所有的交易类型，缺少了非 rmb 协议的交易，因此 2020 年继续启动新的项目来解决这个问题，进一步提升交易树的全面性和准确性。

**智能化根因定位**

2018 年下半年**正式启动了“异常根因定位项目”（简称为“RCA”）。**项目的目标很明确，“快速识别异常并定位根因，期望机器能代替人定位异常或者给出定位的方向”。基本上是希望打破现有的思维模式，由系统或机器人快速识别异常和影响范围，并给出根因分析。

实现思路的图示如下：

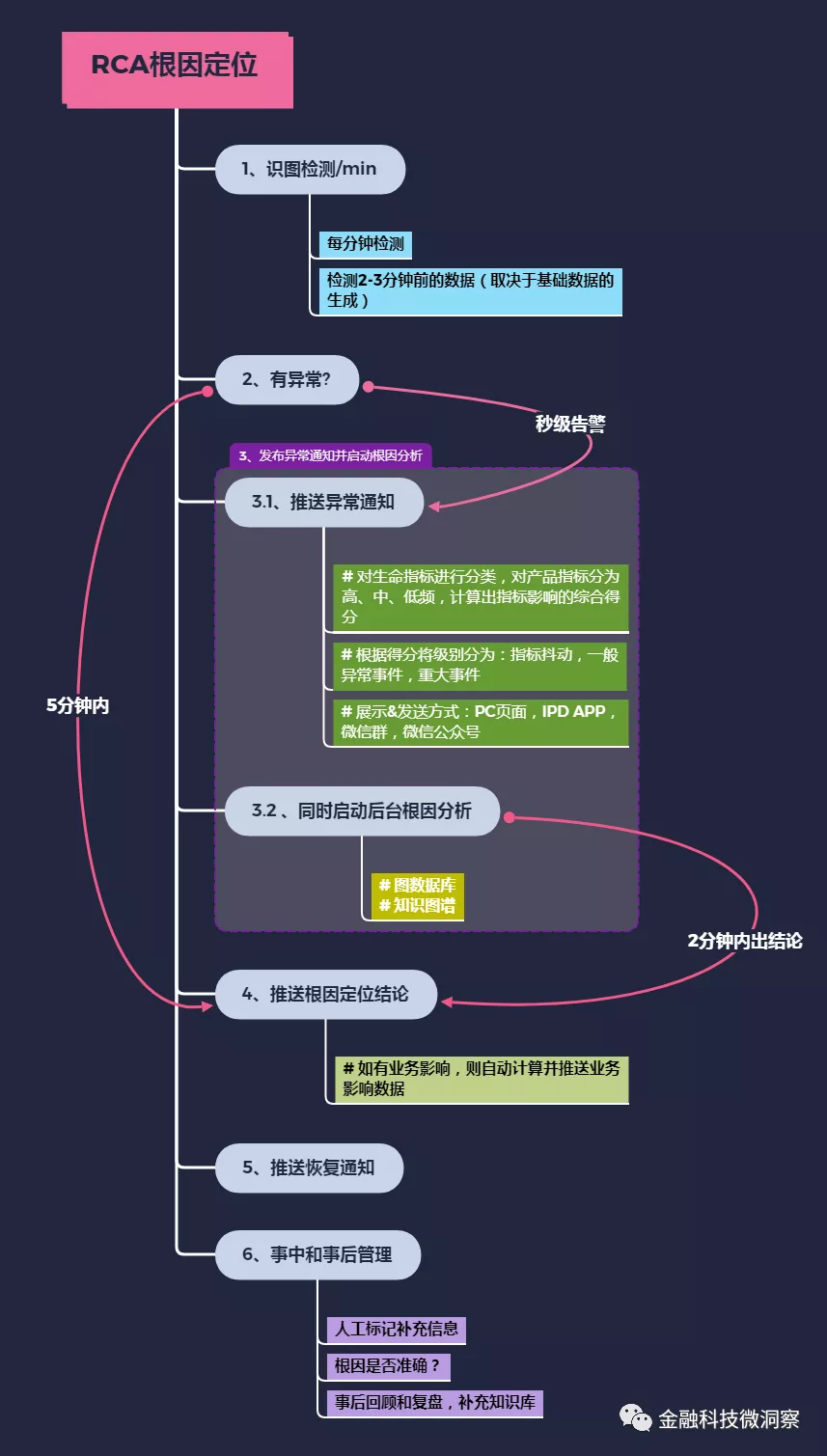


图 1：根因分析方法论图示

图示中的各个事项的解释如下（建议对照上图）：

**1、识图检测 /min：主动监控巡检识别异常**

每分钟对当前生产环境中业务产品黄金生命指标（产品关键功能的交易量，交易时延，系统成功率，业务成功率）进行智能化巡检。

**2、有异常？**

巡检识别当前指标是否偏离检测区间。

这其中遇到的问题如下：

* 算法依赖前期的数据，如果数据规律发生变化，可能引起误报；
* 在交易量低的时段，波动变化大，容易触发告警；

**3.1、推送异常通知**

发生异常后需要推送通知，这其中面临很多细节的问题，需要投入很多精力来管理，通过精细化管理后效果大大提升。上线后当发生异常时，所有人都不需要问影响范围是什么，机器人推送和 PC 页面端，手机移动端都展示的非常清楚。

首先，到底什么样算异常？什么不是异常？一个指标的波动可能是很小的问题，对终端客户也没有影响，如果当大的异常报出来大家受不了。或许大家说可以用 minor，major 和 critical 来定义告警级别即可，为了和 IMS 监控系统的告警区分出来，因此重新定义了一套原则。

将各产品的交易分为高、中、低频，有的产品交易量每天是亿级，有的是百万级，有的是万级，其中的管理模式完全不一样；

再将指标进行分类，例如交易量指标，系统成功率指标重要级别高于平均时延；

根据以上两个维度，对每一笔异常进行打分，不同的分数对应不同的级别。

其次，根据不同的级别，明确不同的通知升级方式。例如目前的告警分为两类：

指标抖动：PC 端页面展示，微信群机器人通知，上报 ITSM 问题单；

一般异常：PC 端页面展示，IPAD 端展示，微信群机器人通知，手机微信企业号展示，公众号通知。事中有事件初定级，过程中根据影响程度有升级机制。事后复盘时再评估事件的真正定级。

**3.2、同时启动后端根因分析**

发送通知同时启动后台根因分析；根因分析是整个项目的难点，怎么做到定位准确？

* 首先数据要足够，前面提到的日志信息，告警信息，接口调用信息，交易链路，CMDB 关系树，变更信息，推广信息，业务行为信息等等；
* 其次是历史异常积累的数据；
* 最后通过多种维度的精细化分析，推导出根因。

根因分析中使用了图数据库、知识图谱等各类前沿技术来支撑根因推导，详细的介绍后续会有细节文章推出。

**4、根据 3.2 的结论推出根因结论：**与 3.1 推送通知一样，推送出根因结论。从推送告警到根因分析在 2 分钟内。

**5、推送恢复通知：**指标正常后推送恢复通知，并计算影响业务量。

**6、事中和事后管理：**

* 异常推送后自动生成 ITSM 事件单跟踪是否需要改进；
* 一线同事协助跟踪和反馈事件原因；
* 运维管理团队同事根据反馈对每一单进行标记，补充知识库；

**收益**

通过内部的灰度，到全面推广上线使用并运行稳定，差不多一年半左右的时间。期间运维团队经历了成功定位到根因时的欢喜雀跃，也经受了连续两个月根因定位准确率持续下降后的挫败感。但总的来说，该项目还是取得了成功，为微众银行的异常管理带来了质的变化，体现在两个方面：

一是在异常时值班群自动获取影响范围，影响交易量以及可能的根因，领导层以及运营方只需要把精力放在如何恢复异常上，无需再花时间去沟通当前异常的内容。

二是在各类指标的体现上：

* 异常升级时间大幅下降（如 2019 年比 2018 年整体提升 60%）
* 异常识别准确率 90% 以上
* 至 2019 年 12 月底，根因准确率 81%

**下一步探索**

通过 2018、2019 年在异常监控和主动定位方面的不断探索，微众银行运维团队实现了智能化运维从无到有的“质的飞跃”，然而，这些成绩的取得仅仅算是开始，未来还有很多挑战亟待解决，运维团队将坚持不懈地进行探索与改进，to boldly go where no one has gone before!