智能运维系列（七）| 化繁为简：业务异常的根因定位方法概述

根因定位分析（RCA）是智能化运维 (AIOPS) 一个重要且难于实现的领域，涉及到归纳分析和演绎推理的相互结合，是从大数定理到逻辑性完备链条推理的综合应用。分布式架构的海量数据为相关分析奠定了基础，但业务异常案例相比于庞大的指标 / 日志数据却显得凤毛麟角，因此需要具备从相关性到因果性的强 AI 能力：基于运维领域知识进行演绎推理，同时因果推导的过程和结论要具有可解释性便于复盘分析和不断优化。

微众银行采用了自主可控的分布式架构，基于大量低成本的 PC server 和开源组件构建复杂的银行系统为亿级客户提供了丰富多样的金融服务。在数量维度，截止 2019 年底，微众银行日金融交易峰值达到了近 6 亿笔。在质量维度，微众银行不断挑战金融系统质量的极限，提出了电信级 99.999% 的全年可用率目标，对故障处理方面对异常定位和恢复时长提出了更高要求，力争实现异常发现后的秒级定位的目标。

分布式架构增大了系统的规模和复杂度，一个业务功能由几十个系统服务组成，软件架构和依赖关系十分复杂，同时这些系统服务由大量的主机 / 容器承载，与众多数据库、中间件及平台组件相关，涉及庞大的基础资源。分布式架构增大了异常定位的难度，依赖人工针对大量数据进行相关分析和跨领域逐一排查的工作方式很难满足定位时效要求，因此运维人员希望有个智能机器人可以在业务异常的时候快速准确的定位问题以便及时恢复业务。

根因定位分析（RCA）是智能化运维 (AIOPS) 一个重要且难于实现的领域，涉及到归纳分析和演绎推理的相互结合，是从大数定理到逻辑性完备链条推理的综合应用。分布式架构的海量数据为相关分析奠定了基础，但业务异常案例相比于庞大的指标 / 日志数据却显得凤毛麟角，因此需要具备从相关性到因果性的强 AI 能力：基于运维领域知识进行演绎推理，同时因果推导的过程和结论要具有可解释性便于复盘分析和不断优化。RCA 的前提是具备了高质量的可观测数据，这些数据包含大家熟知的链路 (Trace)、指标 (Metric) 和日志 (Log)，也包括运维关注的告警（Alarm）、变更（Change）等信息。微众银行的监控系统（IMS）实现了系统应用、平台组件和基础架构的指标监控和告警集中，同时也具备了交易和应用日志的监控能力。智能化平台（KNOWING）实现了业务黄金指标的异常检测以及基于消息总线数据的业务链路绘制。配置管理数据库（CMDB）实现了从基础架构到业务产品的配置项自动管理，提供了实时准确的配置及其关系数据。自动化运维系统（AOMP）实现了应用变更版本发布的统一管理。

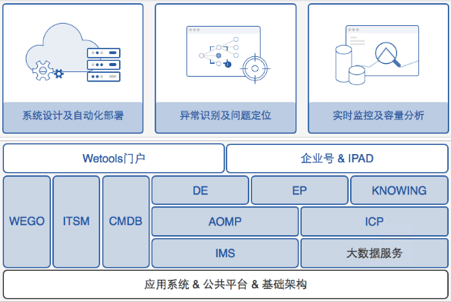


图 1 微众银行 AIOPS 体系

RCA 基于这些系统平台开展建设，分为异常交易分析、异常相关性分析和根因推导等几部分：

**一、异常交易分析**

业务黄金指标是直接反应业务异常且影响客户使用金融服务的指标，我们选定产品入口子系统的关键接口交易统计指标作为业务黄金指标，包括交易量、成功率和耗时等。业务黄金指标异常是由异常交易引起的，准确识别本次事件的异常交易并针对这些交易做全路径排查和关联指标 / 日志的聚类分析可以确定根因方向。其原理类似于控制新冠疫情的过程，即找到确诊病人（异常交易）并进行流行病学追踪（全路径排查）和聚集地分析（聚类分析），从而定位传染源头进而控制扩散（定位故障）。

指标异常时间段内，异常交易判定为下面两种情况：交易返回码标记为失败的交易和高于平日耗时基线的交易。通过交易流水（交易唯一标识）从 KNOWING 获取每笔交易的系统调用路径 (TRACE)，从 IMS 中获取交易路径上每个系统接口的指标数据（METRICS）和日志（LOG）信息，依据下游影响上游的原则并排除常见错误干扰，针对异常交易做聚类分析归纳出导致众多交易链条出错的异常子系统和接口、相关出错信息和聚集 IP，从而确定根因的大致方向。

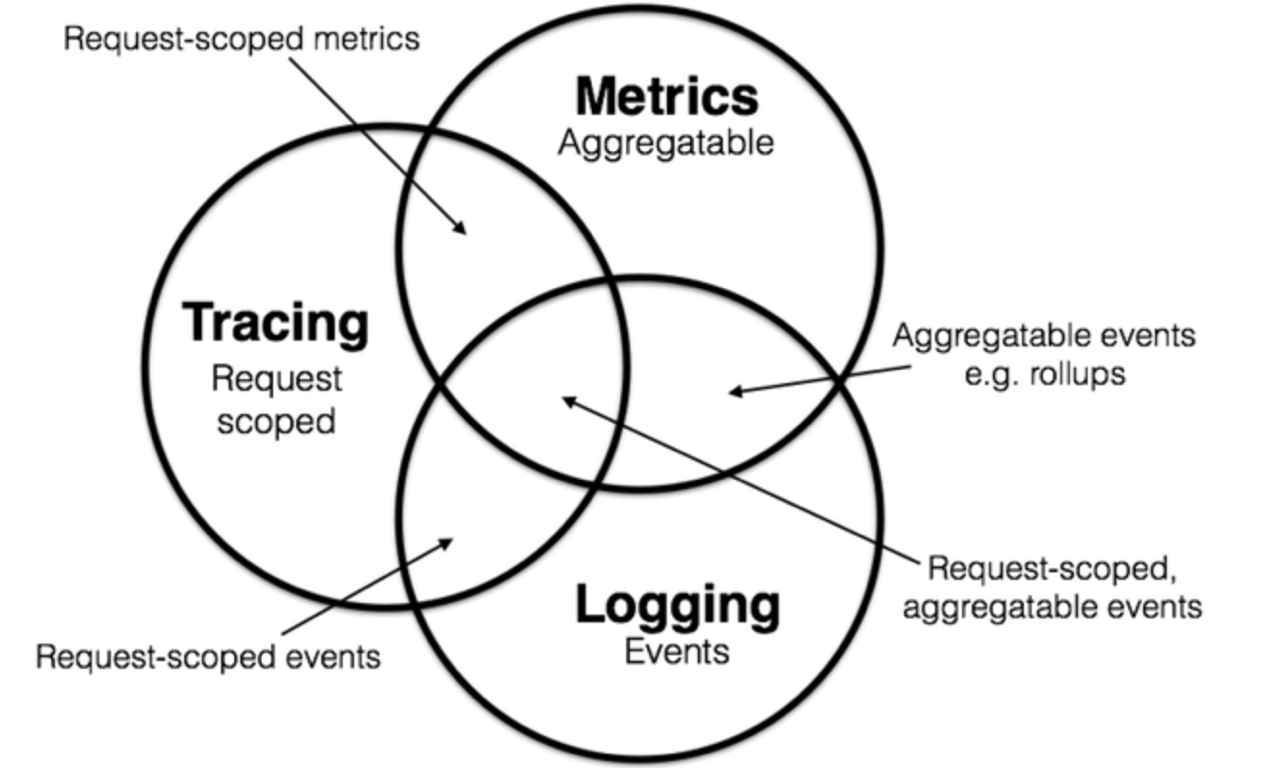


图 2 Metrics, tracing 和 logging 的关系（引自：Peter Bourgon 《Metrics, tracing, and logging》）

**二、异常相关性分析**

针对异常时间段的运维数据通过交易链路路径信息、CMDB 关系数据、历史数据模型等进行相关分析提取有效信息用于最终根因的推导。这些运维数据包括告警（含指标异常）、日志和变更操作等，具体的分析过程如下：

**1. 告警分析**

IMS 日均 2000 条告警，每月不经过收敛的原始告警约 14 万条，每条告警均有压缩策略，如果按照指标越限来统计的告警次数则更加庞大。由于分布式架构的高可用性以及告警定义的差异性，每条告警需要运维人员去处理和关注但不一定会影响业务和黄金指标，因此在业务异常发生时甄别出根因告警显得尤为困难，我们采用了下面一些方法：

（1）影响业务的告警筛选，应用关联分析算法挖掘历史告警的频繁项集并提取关联规则, 归纳引起业务异常的并发告警类型集合。例如主机的磁盘只读会立即引起业务异常，空间使用率越限则不会马上影响业务。

（2）告警相关性分析，结合交易路径信息和 CMDB 关系数据生成本次异常相关的应用实例和基础节点信息，异常窗口新增告警的对象信息与这些数据进行匹配生成相关告警集合。交易路径经过的节点信息认为是强相关，例如异常交易聚集的主机告警；交易路径上未经过通过 CMDB 关联到的节点信息认为是弱相关，例如核心网络、消息总线节点告警。针对弱相关告警需要结合业务影响面、异常范围匹配度、并发特征和历史告警频度等进行加权评分，依据强弱相关告警统一评分结果生成告警根因的概率排名。例如弱相关的核心网络抖动告警很多时候不影响业务，但如果伴随高频产品场景和多个业务集群异常，并发数据库网关不可达、应用超时日志，则抖动告警很可能是根因。

**2. 日志分析**

应用日志是异常分析和问题定位的主要抓手，运维人员尝尝利用异常交易流水上下文的出错日志来定位问题。IMS-LOGM 统计到全行应用系统累计应用日志 291T/ 天、业务交易日志 13T/ 天（截止 2019 年底），这么大量的日志既是宝藏也是挑战。RCA 主要针对关键系统的应用日志进行实时采集和分析。应用 NLP 技术实现文本的聚类分析和异常检测，聚类分析把大量的日志信息进行分类，减少同类日志的干扰；异常检测针对日志中的错误进行主动发现和归类，传统日志监控依赖人工配置错误关键字来识别异常，定义标准和使用目的的不同导致错误种类繁多，而且多数和业务影响不相关。RCA 日志检测采用统一标准对错误日志进行识别和分类，基于历史出现频度进行常见错误的排除，提取 OOM/DB 异常等影响业务的错误类别作为根因分析参考。

**3. 发布变更和业务操作修改**

应用版本发布、基础架构变更、业务操作和配置修改等都会引起业务异常。应用版本和变更信息关联类似于告警关联分析，变更实际操做时间与异常时间相符、发布单元在异常场景的交易路径或者 CMDB 关系链路上；基础架构变更需要结合历史影响面以及并发特征来组合判断；业务操作和配置修改依据时间相关性、产品场景关联性和影响指标匹配性来进行分析。

**三、根因推导**

传统根因推导过程是运维工程师通过对软件架构和调用关系的理解将异常发生时的告警、日志等信息联系在一起，应用运维知识经验来排查推导异常根因，相当于在大脑中存储和训练了一个知识图谱。其中最大的挑战在于，运维工程师的知识经验存在差异而且往往仅精通本领域知识，同时人脑存储的信息量也相对有限。然而，RCA 应用图形数据库（图形数据库是一种非关系型数据库，应用图形理论存储实体之间的关系信息），可以针对每个异常事件创建一个覆盖多应用域及基础架构的全专业图谱，沉淀运维知识进行因果推导。相比于专家规则引擎系统，基于图数据库的知识图谱更利于开发维护，并且具备结合机器学习实现复杂推理和新知识发现的扩展性，可视化的推导链路也具有较好的可解释性易于复盘和优化。

针对异常事件建立的知识图谱包含每笔异常交易的路径信息、CMDB 关联的数据库等基础架构信息、相关性分析得出的告警 / 日志 / 变更信息，针对这些数据基于基础组件影响上层应用等运维知识进行因果推导得出根因，如果存在多种结论则依据加权评分进行可能性排名。

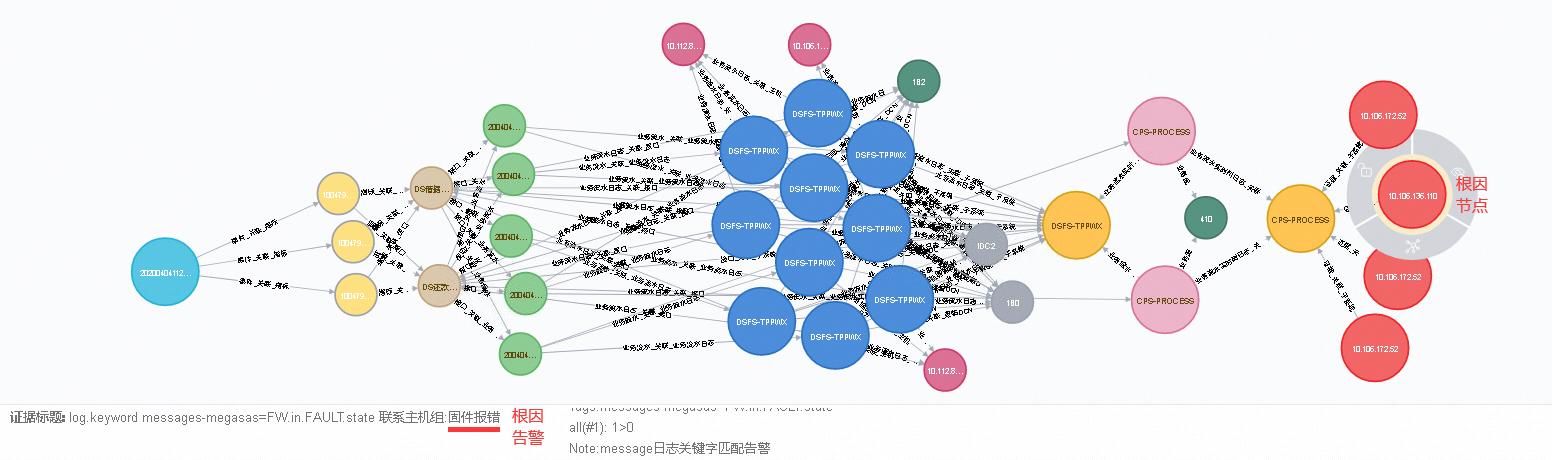


图 3 一个真实案例的图推导

RCA 对基础数据质量有非常高的要求，告警对象信息记录不标准、日志采集的延迟等均会导致最终结果错误。在开展 RCA 建设的时候很多方法也不是一撮而就，无不经过反复验证、不断复盘、持续迭代优化才最终落地建成工程化的应用系统。RCA 涉及到机器学习 / 深度学习、NLP 和知识图谱等众多 AI 技术的综合应用, 但也不能简单的把数据丢给算法来解决问题，例如 RCA 需要业务场景对应的所有接口指标来进行分析定位，曾尝试应用聚类算法在海量指标数据中寻找相似曲线进而确定业务场景关联的同类指标，但结果无法满足要求，最终通过日志中交易流水来关联接口信息的工程化做法确定了相关的指标集合。

微众银行的 RCA 经过前期建设已初显成效，目前定位准确率达到 80% 以上。进一步提升定位准确率以及缩短定位时效到秒级、让 RCA 具有自学习能力是接下来的建设目标，对未来有如下展望：

1. 交易链路是业务异常定位的基石，链路信息越丰富、经过的节点数据越完整则定位问题越容易。结合 APM（应用性能监控系统）的建设丰富链路上的应用程序内部调用信息，补充链路经过的消息总线、域名服务等更多基础节点信息，实现应用代码方法及 SQL 语句级的异常定位。

2. 告警是异常定位的首选参考，实时准确的告警可以让 RCA 事半功倍，一个引起业务异常的重要指标异动如果不能及时检测出来，就要通过大量其他相关辅助信息来弥补论证。针对主机 / 容器、数据库 / 网络等重要基础设施性能指标开展异常检测和无阈值标准化告警，提升基础指标异常发现的实效和准确性。

3. 希望 RCA 具有自学习能力。尝试建立异常信息的全量大规模图谱，包含完整的 CMDB 关系和告警信息，应用图谱技术进行实体连接和相关规则的知识挖掘来不断补充完善现有知识；尝试建立基于向量的知识图谱，通过数值计算与深度学习模型集成来发现新知识和隐性知识。

如果希望了解我们在智能运维中使用的机器学习算法以及支持根因分析的具体方法，请参阅该系列其他文章。