智能运维系列（一）| AIOps 的崛起与实践

**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。

IT 运维其实早在科技公司创建之初即存在了，只是一开始与其说是运维，不如说是网管与电脑维修。公司与电脑有关的都在 IT 运维的工作范围内。随着科技的进步，IT 的岗位划分也越来越细致，才逐渐形成了与现在相似的对业务连续性直接负责的 IT 运维职能部门。而近 10 年随着各种技术浪潮的爆发，IT 运维价值才逐渐显现出来。

通过如下一张图就可以看到运维发展的几个阶段，早期的手工运维，到流程化、标准化运维，再到平台化、自动化运维，最后到近十年的 DevOps（研发运营一体化）和 AIOps（智能运维）。运维发展其实和各种技术更新是密不可分的，也是各种技术重要的实验田。

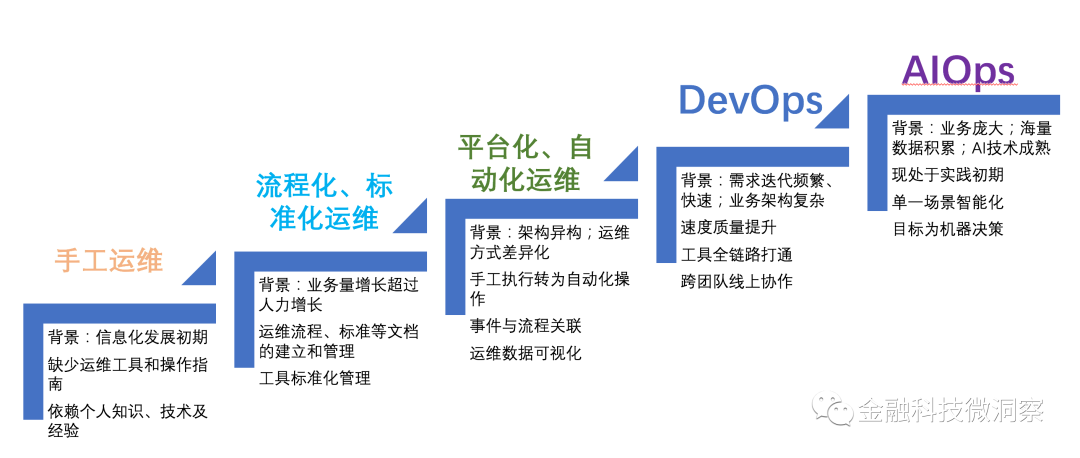


图 1：IT 运维发展示意图

对于企业而言，IT 运维的价值体现在对业务稳定、运行安全和提效降本三个方面的保障与控制。近年来实体业务纷纷发展线上交易，IT 投入逐渐加大，业务的稳定与安全直接影响企业效益，IT 在效率与成本上的平衡也成为各家企业需要面对的问题。

IT 运维行业转向 DevOps 或 AIOps 已经是必然的选择。DevOps 将软件全生命周期的工具全链路打通，结合自动化、跨团队的线上协作能力，实现了快速响应、高质量交付以及持续反馈。AIOps 的落地将为企业运营和运维工作在成本、质量、效率等方面的调整提供重要支撑。

目前 AIOps 的主要应用场景有异常告警、告警收敛、故障分析、趋势预测等。

下面以几个公司的案例为例来分析 AIOps 目前在企业中的实施情况。

**（一）美团业务稳定性保障平台 AIOps 实践**

2019 年 6 月，在全球技术领导力峰会上，来自美团的资深技术专家宋斌从美团的即时物流业务谈起，介绍了美团面临系统稳定性方面极高的挑战：峰值流量高、瞬间峰值大、业务链路长、线上业务复杂度高、故障敏感，影响履约完成率，造成赔付和客诉。经过一年多的努力，从人工运维状态逐步转化为自动化运维，甚至在尝试使用机器学习的能力，提升运维整理效率。

在此次分享中，我们可以看到美团主要在容量评估、弹性设计、故障诊断、风险防范方面逐步建立了全面可靠的自动化可运维的系统。



图 2: 美团配送实践概览

其中在故障诊断方面既包含了故障检测，也包含了根因定位。即回答了故障长什么样，也回答了故障是什么，根因是什么。从研发人员和管理人员两个角度展示故障信息。

在核心功能上不只完成了故障收敛，还完成了链路监控和故障拓扑展示，告警升级等，同时对故障进行相对完整的标记、运营、管理、跟踪等工作。



图 3: 美团故障识别模块

在根因定位思路上即在传统的纵向分析方面进行了根因挖掘，也在调用链调用关系逐层递归的横向分析方面有所创新。



图 4: 美团故障根因分析维度

从结果上看，定位时间和故障覆盖率都很好，极大的降低了运维成本。在整个运维保障体系中，机器学习主要应用在了端到端链路监控、以及失败率监控的动态阀值上（基于 LOF 算法的异常检测），但同时也辅助以规则。



图 5: 美团故障识别收益

**（二）百度 AIOps 在监控报警方面的应用场景**

在 2019 年的全球架构师峰会上，百度资深研发工程师周伟 / 范月林介绍了百度落地 AIOps 过程中在智能异常检测和智能报警合并两个监控场景取得的收获。

如下图是百度的报警系统业务模型，在异常判断阶段，面临的挑战是模型变更频繁，算法不固定。为了解决这个问题，百度实现了算法和模型分离，迭代周期从月缩短到日。验证算法策略的平台分为了离线环境、近线环境、在线环境，使算法在一个可靠的数据环境中验证。

在通告发送阶段，由于可能面临报警风暴的挑战，需要做报警合并。百度从两个方面进行了合并，按部署架构合并和跨部署架构的合并。其中按部署架构合并主要基于配置管理系统的基础关联关系，而跨部署架构合并则依赖上下游关系和离线的策略挖掘。

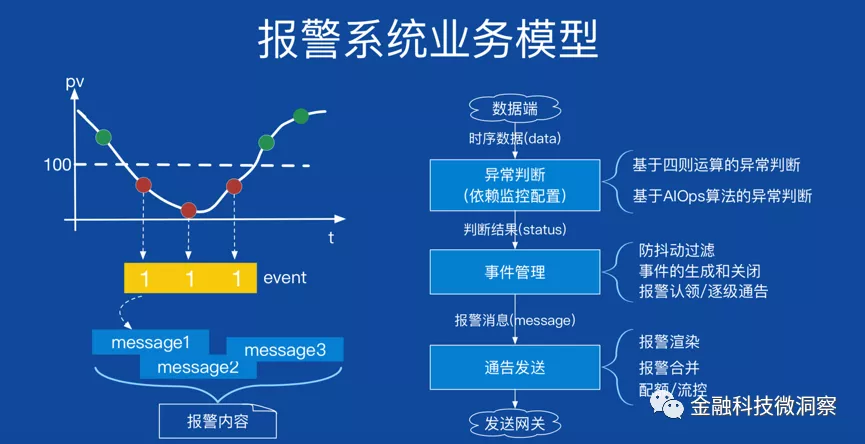


图 6: 百度报警系统业务模型

从效果上看，核心的监控关键指标都有了明显提升。异常检测效果非常好，报警的短信量也大幅降低。



图 7: 百度报警优化收益

**（三）微众银行在异常发现和根因定位方面的 AIOps 实践**

2019 年微众银行在落地 AIOps 过程中，也使用了机器学习算法在多个功能场景中实践，取得了一定的成果。

在异常发现方面，从业务产品可用率出发，定义业务黄金指标（交易量、业务成功率、系统成功率、单位时间交易平均耗时），对黄金指标进行算法结合规则的检测。

在根因定位方面，放弃了传统的组件告警收敛专家经验的定位方法，以消息总线的消息日志为基础，基于算法生成交易链路，异常时绘制故障交易链路图，定位中断子系统，然后以中断子系统为圆心，使用知识图谱算法发散式根因挖掘。另外，历史异常事件的数据也导入知识图谱，修正知识图谱中的专家经验。从运维管理的角度进行异常识别和根因定位标记，不断迭代修订基础模型数据。



图 8: 根因分析逻辑示意图

2019 年微众银行全年异常识别准确率 96%，漏告量低于 1 笔 / 季度。异常通报时常较 2018 年提升 66%。从异常发生开始 5 分钟完成根因定位。异常识别和根因定位的效果都有明显改善。



图 9: 微众银行异常识别与根因定位准确率



图 10: 微众银行异常根因定位收益

从以上比较成功的 3 个案例出发，我们可以尝试总结一下目前 AIOps 在 IT 运维领域实践的几个特点：

(1) 机器学习算法落地场景首推异常指标识别。因为工程化成本低，较容易产生效果。一两个核心开发人员即可。而且一旦成功，运维人员能获得最直接的收益。

(2) 从异常指标出发，衍生到异常追溯原因的各类场景。当识别出了异常，自然希望得到答案。例如以上案例所示美团和微众银行都在根因定位上有所突破，同时都使用到了交易链路，所不同的是，美团是依靠人工管理标注完成，而微众银行使用了算法结合消息总线的日志自动产生。所以，从纯粹数据层面的机器学习逐渐过渡到逻辑层面的机器学习。

(3) 算法结合规则才能达到实际生产运维期望的效果。算法不能解决所有问题。如果要更好的解决实际运维场景的问题，算法结合规则或专家经验才是宝贵的 AIOps 落地经验。

(4) 标准化、数据仓库式的运维数据是基础，大量运维格式化基础数据是机器学习的先决条件。所以，从组织架构到运维工具都应尽量做到集中化管理。

(5) 类似容量预测或事件预测等类似预测类的场景将是 AIOps 的下一个探索方向，主动发现与预测将是 IT 运维未来需要解决的问题。

IT 技术架构从“IOE 架构”逐步走向了“互联网架构”；运维体系从 ITIL 走向 DevOps；运维平台从自动化走向了 AIOps；运维核心从关注平台走向数据资产。智能化以及数据化是未来 IT 运维的总体趋势，互联网业务的连续性保障的方方面面都将在这次 IT 运维变更中发生巨大变化。AIOps 也会在更多运维场景中发挥不可替代的作用。