

Framework de Decisão Causal Adaptativa sob Restrições Dinâmicas para AIOps

José Benedito Brito¹, Aleteia Araujo¹

¹Departamento de Ciências da Computação – Universidade de Brasília (UnB)
Brasília – DF – Brasil

jose.brito@aluno.unb.br, aleteia@unb.br

Abstract. *Modern big data systems generate massive observability data, making manual analysis unfeasible and driving the adoption of Artificial Intelligence for IT Operations (AIOps). This paper presents a high-performance framework designed to mitigate the efficiency-accuracy paradox observed in modern literature. While advanced language models provide accurate root cause analysis, the large majority existing state-of-the-art architectures fail to address strict production latency constraints. The proposed framework implements a layered intelligence architecture using streaming analytics pipelines, small language models, and dynamic lambda-constrained decision engines to enable trustworthy, real-time autonomous remediation in large-scale environments.*

Resumo. *Sistemas modernos de Big Data geram volumes massivos de dados de observabilidade, tornando a análise humana inviável e impulsionando a Inteligência Artificial para Operações de TI (AIOps). Este artigo apresenta um framework de alto desempenho projetado para mitigar o paradoxo eficiência-precisão identificado na literatura recente da área. Embora modelos avançados de linguagem proveem análises de causa raiz acuradas, a grande maioria das abordagens vigentes falham em atender restrições estritas de latência em produção. O framework proposto implementa uma arquitetura de inteligência em camadas utilizando pipelines de processamento de streaming, modelos de linguagem compactos e motores de decisão dinâmica restritos por parâmetros lambda para viabilizar remediação autônoma confiável em tempo real em ambientes de larga escala.*

1. Introdução

O crescimento exponencial da telemetria em arquiteturas de nuvem e ecossistemas distribuídos ultrapassou a capacidade de operação centralizada em analistas humanos. Nesse cenário, as abordagens de *Artificial Intelligence for IT Operations* (AIOps) emergiram como o paradigma essencial para automação de triagem, diagnóstico e autorrecuperação (*self-healing*). Recentemente, a área transitou de modelos preditivos discriminativos clássicos para Grandes Modelos de Linguagem (LLMs) e sistemas multiagentes baseados em raciocínio cognitivo [Joshi 2025].

Contudo, a introdução de modelos generativos massivos impõe um desafio de Processamento de Alto Desempenho (PAD). O alto custo computacional e a latência de inferência dessas arquiteturas frequentemente violam as janelas temporais estritas definidas por Acordos de Nível de Serviço (SLA) no plano de dados [Kapoor and Kas 2026].

Este trabalho aborda esse *trade-off* crítico por meio de duas contribuições integradas: (i) a consolidação empírica de métricas de literatura avaliando os gargalos de tempo de execução na fronteira de AIOps; e (ii) a proposição de um *framework* de decisão causal adaptativa projetado para tomar decisões automatizadas, entender a causa efeito por trás dos dados (causalidade) e ajustar o seu comportamento em tempo real à medida que o ambiente muda (adaptativo), sob restrições dinâmicas, por exemplo de latência e risco.

2. Gargalos Operacionais na Literatura de Inteligência de TI

O mapeamento do estado-da-arte revela o comportamento e a viabilidade operacional de modelos de aprendizado profundo (*Deep Learning*) aplicados a fluxos heterogêneos de logs, métricas e *traces* distribuídos [Bhuyan and Elmroth 2023]. A análise consolidada de implementações em ambientes de larga escala revela quatro lacunas severas que impedem a adoção industrial de AIOps autônomo e resiliente:

1. O Paradoxo Eficiência-Acurácia (Gap de Restrições): A maioria dos sistemas inteligentes operacionais propostos omitem *benchmarks* de tempo de execução ou latência de inferência em seus pipelines reais. Modelos massivos atingem elevada acurácia ($F_1 - score > 0.96$) em diagnósticos desconectados (*offline*), mas geram complexidade assintótica de tokenização de $O(|X_t|^2)$ que inviabiliza respostas em segundos em regime de alta vazão ($\Lambda_t \geq 10^6 \text{ eventos/s}$);
2. A Ilusão da Autonomia (Gap de Tarefas): Embora abordagens recentes baseadas em grafos e multiagentes [Wang et al. 2025, Zhang et al. 2026] aleguem suporte à automação, poucas implementações abordam efetivamente a fase de execução ou remediação autônoma. A grande maioria restringe-se a regimes consultivos (*advisory*), gerando alertas não-determinísticos sem intervenção direta sobre o ambiente;
3. Riscos de Alucinação (Gap de Modelos): A taxa de alucinações estocásticas em LLMs aplicados diretamente a infraestruturas de TI alcança marcas críticas [Kapoor and Kas 2026], introduzindo o risco de comandos erráticos capazes de desencadear falhas em cascata no plano de rede e armazenamento;
4. Dependência de Dados Legados (Gap de Dados): A maioria das validações experimentais utilizam bases estáticas e obsoletas, desconsiderando o desvio de conceito (*concept drift*) e a volatilidade de cargas de trabalho dinâmicas reais.

3. O Framework Proposto

Para mitigar os gargalos diagnosticados, definimos um framework estruturado com uma arquitetura de inteligência em camadas [Kephart and Chess 2003]. O cerne da abordagem reside no desacoplamento entre a detecção em tempo real de baixa latência e o planejamento cognitivo centralizado de alta granularidade, utilizando um motor de otimização parametrizado por restrições λ .

As camadas operacionais dividem-se em funções específicas:

- Camada de Observabilidade Multimodal: Executa a ingestão contínua de fluxos via coletores distribuídos e instrumentação. Constrói dinamicamente um Grafo do Sistema acoplando topologias heterogêneas do ecossistema distribuído;
- Camada de Detecção e Modelagem Causal: Utiliza algoritmos de processamento em fluxo de baixa latência para identificar anomalias no plano de dados. Aplica

Inferência Causal baseada em Modelos Causais Estruturais (SCM) para gerar o subgrafo explicativo da anomalia;

- Camada de Estimação de Confiança: Qualifica quantitativamente a incerteza do diagnóstico antes de qualquer tomada de ação, avaliando a consistência das métricas;
- Camada Núcleo de Decisão: Executa uma otimização restrita na qual o parâmetro λ atua como o peso dinâmico das restrições. Se o ambiente encontra-se sob estresse severo ou criticidade de SLA, o atualizador adaptativo eleva o valor de λ , tornando o motor de decisão altamente conservador;
- Camada de Gating Cognitivo e Execução: Implementa o conceito de autonomia graduada. Ações cujo diagnóstico possua alta confiança, baixa incerteza e validação por salvaguardas semânticas determinísticas (*policy-as-code*) são despachadas para execução autônoma via APIs. Caso contrário, o framework degrada para os modos assistivo ou *human-in-the-loop*;
- Camadas de Realimentação e Feedback: Avaliam o impacto pós-intervenção, após uma janela Δt , para reajuste de λ ;
- Camada de Governança e Aprendizado: Aprende a partir do histórico de decisões e resultados. Gerencia o ciclo de vida dos modelos (*drift detection*), implementa suporte à explicabilidade (XAI).

4. Ambiente em Implementação e Resultados Preliminares

Para a validação experimental, o ambiente de homologação utiliza um *cluster* Hadoop sob virtualização *Proxmox VE* e as camadas do *framework* foram instanciadas em *containers* Docker. O plano de dados coleta informações do ecossistema de forma desacoplada e assíncrona por meio do OpenTelemetry. Estamos utilizando o Apache Kafka como barramento de dados e o *Neo4J* para grafos de topologia. O status atual do desenvolvimento cobre a ingestão em fluxo e a detecção de anomalias em tempo real por meio do algoritmo Half-Space Trees (HST). Este ambiente serve como banco de testes para a próxima fase do trabalho, projetada para calibrar o atualizador de λ e avaliar o *framework* fim-a-fim.

Referências

- Bhuyan, J. M. and Elmroth, E. (2023). Anomaly detection and resolution on the edge: Solutions and future directions. *Proceedings of the 17th IEEE International Conference on Service-Oriented System Engineering (SOSE)*.
- Joshi, S. (2025). Llmops, agentops, and mlops for generative ai: A comprehensive review. *International Journal of Computer Applications Technology and Research*.
- Kapoor, R. and Kas, M. (2026). Breaking the observability tax: Dynamic resolution anomaly detection via topology-aware active llm agents. *IEEE Access*.
- Kephart, J. O. and Chess, D. M. (2003). The vision of autonomic computing. *Computer*, vol. 36, no. 1,.
- Wang, S., Zhao, Y., Meng, W., and Pei, D. (2025). GALA: Can graph-augmented large language model agentic workflows elevate root cause analysis? *arXiv preprint arXiv:2508.12472*.
- Zhang, J., Liu, X., Wang, Z., and Huang, K. (2026). Root cause analysis method based on large language models. *arXiv preprint arXiv:2602.08804*.