

Monitoramento Granular de Clusters Cassandra sob Cargas Dinâmicas para Predição de Latência

Maria Fernanda N. Gouveia¹, Johny M. B. Ribeiro¹, Paulo R. S. L. Coelho¹,
Luis F. Faina¹, Rafael Pasquini¹

¹Faculdade de Computação (FACOM) – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brasil

{maria.gouveia, johny.ribeiro, paulocoelho}@ufu.br

{luisfaina, rafael.pasquini}@ufu.br

Abstract. *This work presents a virtualized experimental environment based on Proxmox and Apache Cassandra for monitoring distributed services under dynamic workloads. A monitoring pipeline with one-second granularity was developed to collect infrastructure metrics and tail latency measurements from Cassandra read and write operations. The collected dataset was used to train a Random Forest regressor for predicting the 95 percentile latency under varying workload conditions. Experimental results, analyzed through CDFs and prediction metrics, demonstrate that high-granularity monitoring is effective for capturing workload transients and supporting latency prediction.*

Resumo. *Este trabalho apresenta um ambiente experimental virtualizado baseado em Proxmox e Apache Cassandra para monitoramento de serviços distribuídos sob cargas dinâmicas. Foi desenvolvido um pipeline de monitoramento com granularidade de um segundo para coletar métricas de infraestrutura e medições de latência de cauda das operações de leitura e escrita do Cassandra. O conjunto de dados coletado foi utilizado no treinamento de um modelo Random Forest para predição do 95 percentil de latência sob diferentes condições de carga. Os resultados experimentais, analisados por meio de FDAs e métricas de predição, demonstram que a coleta em alta granularidade é eficaz para capturar transientes de carga e apoiar a predição de latência.*

1. Introdução

As arquiteturas de redes de próxima geração, como o 6G, propõem a integração nativa de Inteligência Artificial (IA) para automação, otimização de recursos e detecção de falhas [ITU-R 2023, Scalco-Neto et al. 2021]. Nesse contexto, a observabilidade da infraestrutura torna-se fundamental para fornecer métricas capazes de alimentar modelos preditivos e apoiar mecanismos autônomos de gerenciamento.

Este trabalho utiliza um ambiente virtualizado baseado no hipervisor Proxmox [Proxmox Server Solutions 2024], tendo o Apache Cassandra [Lakshman and Malik 2010] como serviço distribuído alvo. O ambiente experimental emprega um gerador de carga dinâmica capaz de produzir diferentes condições de operação, permitindo a coleta de métricas sob distintos regimes de estresse.

O principal diferencial do trabalho está no desenvolvimento de um *pipeline* de monitoramento com granularidade de um segundo, responsável pela coleta de métricas de infraestrutura e estatísticas de latência das operações de leitura e escrita do Cassandra. O sistema consolida automaticamente os dados monitorados (variáveis X) e os rótulos associados às métricas de desempenho da aplicação (alvo Y), permitindo a construção de um *dataset* voltado ao treinamento de modelos de aprendizado de máquina.

Como estudo de caso, foi utilizado um modelo de regressão *Random Forest* [Hastie et al. 2009] para predição da latência de cauda (95 percentil) sob cargas dinâmicas. Os resultados foram avaliados por meio de métricas de predição e estatísticas, como a Função de Distribuição Acumulada (FDA), demonstrando a importância da coleta em alta granularidade para capturar transientes de carga em ambientes distribuídos.

2. Trabalhos Relacionados e Fundamentação

Ambientes virtualizados e distribuídos apresentam elevada variabilidade operacional, exigindo mecanismos eficientes de monitoramento e coleta de métricas para análise de desempenho e automação de recursos. Nesse contexto, aplicações distribuídas executadas sobre infraestruturas compartilhadas podem apresentar oscilações significativas de desempenho em função das cargas de trabalho e do consumo de recursos computacionais.

A literatura aponta que a gestão manual de recursos em infraestruturas virtualizadas torna-se limitada diante da volatilidade das cargas de trabalho [Mijumbi et al. 2016, Al-Khalil 2026, Letaief et al. 2019]. Além disso, especificações como o NWDAF (*Network Data Analytics Function*) incorporam funções de coleta e análise para suporte à predição de carga e detecção de anomalias [3GPP 2023].

Diferente de abordagens tradicionais com baixa frequência de coleta, este trabalho utiliza monitoramento com granularidade de um segundo para correlacionar métricas de infraestrutura e latência de cauda em um *cluster* Apache Cassandra submetido a cargas dinâmicas. Os dados coletados são utilizados no treinamento de um modelo *Random Forest* [Hastie et al. 2009], explorando a capacidade desse algoritmo em modelar relações não lineares entre o estado da infraestrutura e o desempenho da aplicação.

3. Cenário Experimental e Metodologia

O ambiente experimental foi desenvolvido para reproduzir uma infraestrutura virtualizada voltada a serviços distribuídos. A plataforma utiliza o hipervisor Proxmox VE [Proxmox Server Solutions 2024], sobre o qual foi instanciado um *cluster* Apache Cassandra composto por cinco nós virtualizados, cada um com 16 GB de RAM, 4 núcleos, 40 GB de disco e Ubuntu 22.04. Além do *cluster*, o ambiente possui duas instâncias dedicadas: uma responsável pela geração de carga e outra pelo monitoramento das operações de leitura e escrita.

A geração de carga foi realizada com a ferramenta *cassandra-stress*, configurada para produzir cargas dinâmicas de leitura e escrita seguindo um comportamento senoidal ao longo do experimento. Essa estratégia permitiu submeter o *cluster* a diferentes regimes de estresse de maneira controlada, produzindo variações temporais nas métricas de infraestrutura e desempenho.

O monitoramento dos tempos de resposta foi realizado por uma instância isolada executando o *cassandra-stress*. Essa instância coleta estatísticas de latência, incluindo

valores médios e percentis das operações de leitura e escrita. As métricas de latência obtidas foram utilizadas como rótulos (Y), enquanto as métricas de infraestrutura compuseram o conjunto de variáveis de entrada (X).

A coleta de métricas de infraestrutura foi realizada utilizando Prometheus [Prometheus Authors 2026], Grafana [Grafana Labs 2026] e cAdvisor [Google 2024], permitindo monitorar consumo de CPU, memória e I/O de disco dos *containers*. O principal desafio técnico do trabalho consistiu na redução da granularidade de coleta. Enquanto exportadores tradicionais operam com intervalos superiores a 15 segundos, este trabalho implementou um mecanismo de *scraping* customizado em Python para coleta em intervalos de um segundo.

O sistema de coleta utiliza expressões regulares (*regex*) para extrair métricas diretamente dos *endpoints* do Prometheus, incluindo estatísticas associadas ao 95 percentil das operações do Cassandra. Os dados coletados foram consolidados em um *dataset* CSV contendo métricas de infraestrutura, carga aplicada e estatísticas de latência. Posteriormente, o conjunto de dados foi utilizado no treinamento de um modelo de regressão baseado em *Random Forest*.

4. Resultados e Discussão

O ambiente experimental permitiu a coleta contínua de métricas de infraestrutura e desempenho com granularidade de um segundo, possibilitando observar a dinâmica temporal do *cluster* Cassandra sob diferentes condições de carga. A utilização de um gerador senoidal de carga produziu variações significativas nos tempos de resposta das operações de leitura e escrita, permitindo que o *dataset* consolidado representasse desde períodos de baixa utilização até situações de maior estresse computacional.

A alta frequência de coleta mostrou-se importante para capturar transientes de desempenho que seriam parcialmente omitidos em abordagens tradicionais com janelas temporais maiores. Como consequência, o *dataset* produzido apresenta elevada variabilidade estatística nos tempos de resposta monitorados, incluindo valores reduzidos em períodos de baixa carga e picos de latência associados aos momentos de maior utilização do ambiente virtualizado.

Além da variabilidade temporal, o ambiente experimental favorece a geração de dados com características não-independentes e identicamente distribuídas (não-IID), uma vez que diferentes regimes de operação impactam diretamente o comportamento das métricas monitoradas. Essa característica é particularmente relevante para pesquisas envolvendo aprendizado distribuído e aprendizado federado em cenários de monitoramento de infraestrutura. Mais detalhes sobre a caracterização não-IID do ambiente podem ser encontrados em [Ribeiro et al. 2026].

A Figura 1 apresenta as FDAs referentes ao 95 percentil das latências das operações de leitura e escrita efetivamente observadas e estimadas pelo *Random Forest*. Observa-se a presença de uma distribuição com cauda longa, evidenciando a ocorrência de períodos de degradação de desempenho associados às variações dinâmicas de carga impostas ao *cluster* Cassandra. Esse comportamento demonstra que o ambiente experimental produz condições realísticas de operação, importantes na construção de *datasets* representativos para aplicações de IA.

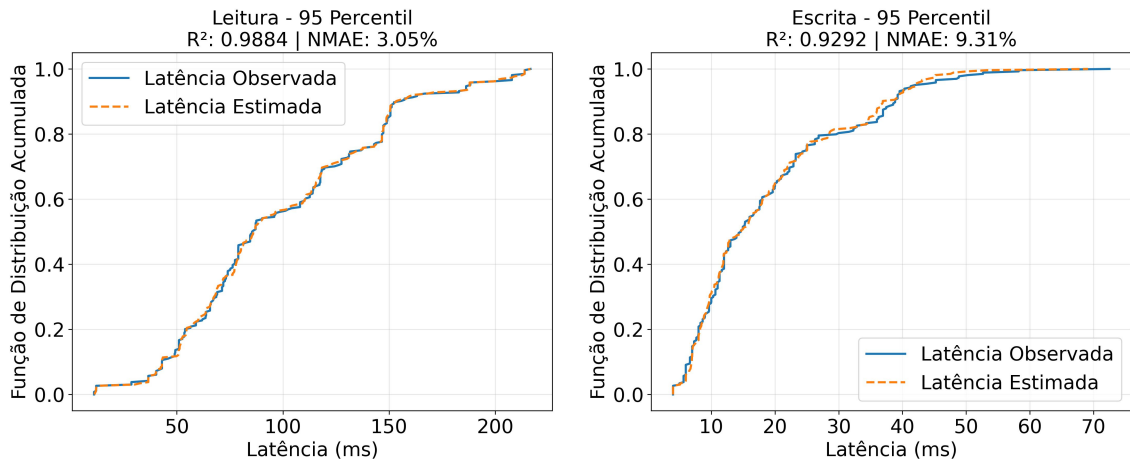


Figure 1. FDA do 95 percentil da latência de operações de leitura e escrita no cluster Cassandra sob cargas dinâmicas.

Com o *dataset* consolidado, foi treinado um modelo de regressão baseado em *Random Forest* [Hastie et al. 2009] para predição da latência de cauda a partir das métricas de infraestrutura coletadas. O objetivo consiste em avaliar a capacidade do modelo em aprender relações não lineares entre o estado do ambiente virtualizado e os tempos de resposta observados no Cassandra. A Tabela 1 apresenta os principais parâmetros utilizados no treinamento do modelo, bem como as métricas de avaliação obtidas nos experimentos.

Table 1. Resultados do modelo *Random Forest* para predição da latência de cauda das operações de leitura e escrita considerando o 95 percentil.

Parâmetro/Métrica	Valor
Alvo predito	95 percentil (Leitura/Escrita)
Número de árvores	50
Profundidade máxima	Sem limite definido
Semente aleatória	42
Divisão treino/teste	70 % / 30 %
R^2	Leitura = 0.9884 / Escrita = 0.9292
NMAE	Leitura = 3,05 % / Escrita = 9,31 %

Os resultados obtidos indicam o potencial da coleta em alta granularidade para aplicações de monitoramento inteligente e predição de desempenho em ambientes distribuídos. A correlação entre métricas de infraestrutura e latência observada nos experimentos reforça a viabilidade da utilização do *dataset* em pesquisas futuras envolvendo modelos de IA aplicados à automação e observabilidade de infraestruturas.

5. Conclusão

Este trabalho apresentou um ambiente experimental virtualizado para monitoramento de *clusters* Apache Cassandra sob cargas dinâmicas, utilizando coleta de métricas com granularidade de um segundo. O ambiente desenvolvido permitiu correlacionar métricas de infraestrutura e latência de cauda em diferentes regimes de operação, produzindo um *dataset* representativo para aplicações de aprendizado de máquina.

Os resultados demonstraram que a coleta em alta granularidade é capaz de capturar transientes de desempenho relevantes em ambientes distribuídos, incluindo períodos de elevada variabilidade temporal nos tempos de resposta. Além disso, o modelo *Random Forest* apresentou capacidade de predição da latência de cauda a partir das métricas monitoradas, indicando o potencial do ambiente para pesquisas envolvendo observabilidade inteligente e análise preditiva. Como continuidade, pretende-se explorar a utilização do *dataset* em cenários de aprendizado distribuído, aprendizado federado e detecção de anomalias em infraestruturas computacionais virtualizadas.

References

- 3GPP (2023). Architecture enhancements for 5g system (5gs) to support network data analytics services. Technical Report TS 23.288, 3rd Generation Partnership Project (3GPP). Version 18.2.0.
- Al-Khalil, A. (2026). Artificial intelligence-driven network slicing: A comparative study of 5g and 6g automation capabilities. *ARO-The Scientific Journal of Koya University*, 14:212–221.
- Google (2024). cAdvisor (Container Advisor). GitHub repository.
- Grafana Labs (2026). Grafana: The open observability platform. Accessed in: 2026-05-10.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition.
- ITU-R (2023). Framework and overall objectives of the future development of imt for 2030 and beyond. Technical Report Recommendation ITU-R M.2160-0, International Telecommunication Union.
- Lakshman, A. and Malik, P. (2010). Cassandra: A decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, 44(2):35–40.
- Letaief, K. B., Chen, W., Shi, Y., Zhang, J., and Zhang, Y.-J. A. (2019). The roadmap to 6g: Ai empowered wireless networks. *arXiv preprint arXiv:1904.11686*.
- Mijumbi, R., Serrat, J., Gorricho, J.-L., Bouten, N., De Turck, F., and Boutaba, R. (2016). Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys & Tutorials*, 18(1):236–262.
- Prometheus Authors (2026). Prometheus: An open-source monitoring system with a dimensional data model. Accessed in: 2026-05-10.
- Proxmox Server Solutions (2024). *Proxmox Virtual Environment Administration Guide*. Vienna, Austria. Release 8.1.
- Ribeiro, J., Vandikas, K., Marquezini, M. V., Rothenberg, C., and Pasquini, R. (2026). An Experimental Framework for Studying Non-IID Data in Federated Learning for Network Telemetry. In *Anais do XLIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 127–140, Porto Alegre, RS, Brasil. SBC.
- Scalco-Neto, H., Araujo, A., Both, C., Oliveira-Jr, A., and Cardoso, K. (2021). Uma visão de arquitetura para redes 6g. In *Anais do I Workshop de Redes 6G*, pages 31–36, Porto Alegre, RS, Brasil. SBC.