

**ANALISIS KOMPARASI PERFORMA LAYANAN *SERVER-BASED* DAN
SERVERLESS DI GOOGLE CLOUD PLATFORM (GCP) (STUDI KASUS:
DEPLOYMENT MODEL MACHINE LEARNING)**

SKRIPSI

Diajukan untuk memenuhi sebagian syarat untuk memperoleh gelar Sarjana
Teknik Program Studi Sistem Telekomunikasi



Oleh
Vina Fujiyanti
2005751

**PROGRAM STUDI S1 SISTEM TELEKOMUNIKASI
KAMPUS UPI DI PURWAKARTA
UNIVERSITAS PENDIDIKAN INDONESIA
2024**

**ANALISIS KOMPARASI PERFORMA LAYANAN *SERVER-BASED* DAN
SERVERLESS DI GOOGLE CLOUD PLATFORM (GCP) (STUDI KASUS:
DEPLOYMENT MODEL MACHINE LEARNING)**

Oleh
Vina Fujiyanti

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Teknik pada Program Studi Sistem Telekomunikasi

© Vina Fujiyanti 2024
Universitas Pendidikan Indonesia
2024

Hak cipta dilindungi undang-undang
Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian,
dengan dicetak ulang, difotokopi, atau cara lainnya tanpa izin dari penulis

LEMBAR PENGESAHAN

VINA FUJIYANTI

**ANALISIS KOMPARASI PERFORMA LAYANAN *SERVER-BASED* DAN
SERVERLESS DI GOOGLE CLOUD PLATFORM (GCP) (STUDI KASUS:
DEPLOYMENT MODEL MACHINE LEARNING)**

Disetujui dan Disahkan Oleh Pembimbing :

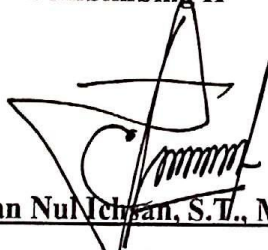
Pembimbing I



Galura Muhammad Suranegara, S.Pd., M.T.

NIP. 920190219920111101

Pembimbing II

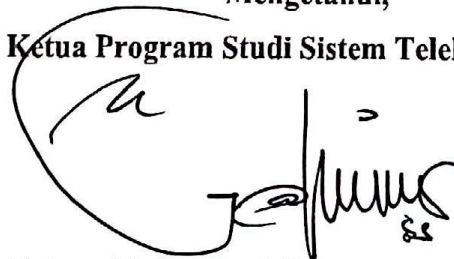


Ichwan Nulichwan, S.T., M.T.

NIP. 920200419900330101

Mengetahui,

Ketua Program Studi Sistem Telekomunikasi



Galura Muhammad Suranegara, S.Pd., M.T.

NIP. 920190219920111101

PERNYATAAN ANTI PLAGIARISME

Dengan ini, penulis menyatakan bahwa skripsi dengan judul “**Analisis Komparasi Performa Layanan *Server-Based* dan *Serverless* di Google Cloud Platform (GCP) (Studi Kasus: *Deployment Model Machine Learning*)**” ini beserta seluruh isinya adalah benar-benar karya penulis sendiri. Penulis tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, penulis siap menanggung sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya penulis.

Purwakarta, Juni 2024

Yang Membuat Pernyataan,



Vina Fujiyanti

NIM. 2005751

UCAPAN TERIMA KASIH

Segala puji dan syukur tak lupa penulis panjatkan ke hadirat Allah Subhanahu Wata'ala. Karena atas berkat rahmat dan hidayat-Nya penulis dapat menyelesaikan penyusunan skripsi ini dengan baik. Penulisan skripsi ini dimaksudkan untuk memenuhi persyaratan dalam memperoleh gelar Sarjana Teknik pada Program Studi Sistem Telekomunikasi.

Penyusunan skripsi ini tentunya tidak terlepas berkat motivasi, dorongan, dan bantuan dari berbagai pihak. Oleh karena itu, penulis mengucapkan banyak terima kasih kepada semua pihak yang telah membantu, di antaranya:

1. Kedua orang tua penulis yang selalu memberikan doa, motivasi, dukungan moral dan finansial sehingga penulis dapat menginjak langkah akhir pendidikan pada perguruan tinggi ini dengan lancar dan baik.
2. Bapak Galura Muhammad Suranegara, S.Pd., M.T. dan Bapak Ichwan Nul Ichsan, S.T., M.T. selaku dosen pembimbing yang telah meluangkan waktu, tenaga, dan pikiran untuk senantiasa membimbing, memberikan masukan dan arahan kepada penulis dalam menyelesaikan skripsi.
3. Tim, pemateri, dan mentor Bangkit Academy yang sudah memfasilitasi penulis untuk mempelajari berbagai ilmu terutama mengenai *Cloud Computing* sehingga penulis dapat mengimplementasikannya dalam penyusunan skripsi ini.
4. *Sobat Villa Rudi*: Anggie, Atik, Alip, Devi, Eci, Galuh, Hanoi, Kagit, Kania, dan *Mbak Nia* yang telah kebersamai, membantu dan memberikan dukungan selama perkuliahan berlangsung.
5. Makna A'raaf Kautsar yang telah banyak membantu dan memberi dukungan selama pengerjaan skripsi ini berlangsung.
6. Semua pihak yang telah kebersamai, mendukung, dan membantu penulis dalam menyelesaikan skripsi ini yang tidak bisa penulis sebutkan satu persatu.

Semoga kebaikan dari berbagai pihak yang telah membantu penulis selama kegiatan perkuliahan berlangsung, termasuk dalam menyelesaikan skripsi ini diberikan balasan oleh Allah Subhanahu Wata'ala. Semoga penulisan skripsi ini bermanfaat, khususnya bagi penulis dan para pembaca secara umum.

ABSTRAK

Penggunaan teknologi *cloud computing* menjadi pertimbangan para pengguna karena biayanya yang efisien, pengelolaan yang terpusat, fleksibel, dan layanannya yang bervariasi, salah satunya layanan komputasi. Penyedia infrastruktur *cloud* seperti Google Cloud Platform (GCP) menyediakan layanan komputasi *server-based* dan *serverless* untuk *deployment* aplikasi seperti model *machine learning*. Baik layanan *server-based* dan *serverless* memiliki karakteristik dan kelebihan masing-masing yang membuat pengguna kesulitan dalam memilih layanan yang sesuai dengan kebutuhan. Oleh karena itu, penelitian ini dilakukan untuk mengkomparasi layanan *server-based* dan *serverless* dengan melakukan *deployment* model *machine learning* untuk mengetahui layanan terbaik berdasarkan pengukuran performa layanan. Performa layanan yang akan diukur pada penelitian ini adalah CPU dan *memory utilization*, *latency*, *pricing*, dan parameter tambahan berdasarkan *developer experiences* yang meliputi kompatibilitas *framework machine learning*, tingkat kemudahan penerapan, dan tingkat ketersediaan dokumentasi. Pengujian dilakukan dengan mengirim 100 kali permintaan terhadap *endpoint* hasil *deployment* melalui HTTP request GET pada JMeter. Hasil penelitian menunjukkan bahwa layanan Cloud Run terbukti ringan dan unggul untuk melakukan *deployment* berdasarkan pengukuran performa karena memiliki utilisasi *resource* yang rendah yaitu CPU *utilization* sebesar 0,05%, *memory utilization* sebesar 0,91%, *latency* sebesar 2,70 ms, serta biaya yang efisien sebesar Rp 6.493. Penelitian ini menunjukkan bahwa layanan *serverless* memiliki performa yang lebih baik dan optimal untuk melakukan *deployment*, khususnya pada model *machine learning*.

Kata kunci: GCP, *server-based*, *serverless*, *deployment*, *machine learning*

ABSTRACT

The use of cloud computing technology is considered by users because of its cost efficiency, centralized management, flexibility, and variety of services, including computing services. Cloud infrastructure providers such as Google Cloud Platform (GCP) provide server-based and serverless computing services for application deployment such as machine learning models. Both server-based and serverless services have their own characteristics and advantages that make it difficult for users to choose the service that suits their needs. Therefore, this research is conducted to compare server-based and serverless services by deploying machine learning models to determine the best service based on service performance measurements. The service performance that will be measured in this research is CPU and memory utilization, latency, pricing, and additional parameters based on developer experiences which include machine learning framework compatibility, ease of implementation, and level of documentation availability. Testing is done by sending 100 requests to the endpoint through HTTP request GET on JMeter. The results showed that the Cloud Run service proved to be lightweight and superior for deployment based on performance measurements because it has low resource utilization, namely CPU utilization of 0.05%, memory utilization of 0.91%, latency of 2.70ms, and an efficient cost of Rp 6,493. This research shows that serverless services have better and optimal performance for deployment, especially in machine learning models.

Keywords: *GCP, server-based, serverless, deployment, machine learning*

DAFTAR ISI

LEMBAR PENGESAHAN	i
PERNYATAAN ANTI PLAGIARISME	ii
UCAPAN TERIMA KASIH	iii
ABSTRAK	iv
ABSTRACT	v
DAFTAR ISI.....	vi
DAFTAR GAMBAR.....	x
DAFTAR TABEL	xii
DAFTAR LAMPIRAN.....	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang Penelitian	1
1.2 Rumusan Masalah Penelitian	3
1.3 Tujuan Penelitian.....	3
1.4 Batasan Masalah Penelitian.....	3
1.5 Manfaat Penelitian.....	4
1.6 Sistematika Penulisan.....	5
BAB II KAJIAN PUSTAKA	6
2.1 <i>Cloud Computing</i>	6
2.2 Layanan <i>Server-Based</i> dan <i>Serverless Computing</i>	7
2.2.1 Layanan <i>Server-Based</i>	7
2.2.2 Layanan <i>Serverless</i>	8
2.3 Google Cloud Platform	8
2.3.1 Google Compute Engine	9
2.3.2 Google Kubernetes Engine	11
2.3.3 Google App Engine.....	13
2.3.4 Cloud Run	14
2.3.5 Cloud Functions	15
2.3.6 Vertex AI.....	16
2.4 <i>Machine Learning</i>	16
2.5 Performa Layanan	18
2.5.1 CPU dan <i>Memory Utilization</i>	18

2.5.2	<i>Latency</i>	19
2.5.3	<i>Cost Per Transactions</i>	20
2.6	<i>Performance Testing</i>	20
2.7	Apache JMeter	20
2.8	Google Cloud Monitoring	22
2.9	Penelitian Relevan	23
BAB III METODOLOGI PENELITIAN		29
3.1	Jenis Penelitian	29
3.2	Alur Penelitian	29
3.2.1	Studi Literatur	31
3.2.2	Identifikasi Masalah	31
3.2.3	Perumusan Masalah	31
3.2.4	Perancangan Desain Penelitian	31
3.2.4.1	Kebutuhan Penelitian	31
3.2.4.2	Perancangan Arsitektur Sistem	32
3.2.4.3	Konfigurasi Sistem	35
3.2.5	Pelaksanaan Penelitian	36
3.2.5.1	<i>Deployment Layanan Server-Based</i>	36
3.2.5.2	<i>Deployment Layanan Serverless</i>	37
3.2.5.3	Pengumpulan Data Performa	38
3.2.6	Analisis Data	40
3.2.7	Penarikan Kesimpulan	40
BAB IV HASIL DAN PEMBAHASAN		41
4.1	<i>CPU Utilization</i>	41
4.1.1	<i>Layanan Server-Based</i>	41
4.1.2	<i>Layanan Serverless</i>	44
4.1.2.1	Kubernetes Engine	44
4.1.2.2	App Engine	45
4.1.2.3	Cloud Run	47
4.1.2.4	Cloud Functions	48
4.1.2.5	Vertex AI	50
4.1.3	Komparasi Performa <i>CPU Utilization</i>	51

4.2	<i>Memory Utilization</i>	52
4.2.1	Layanan <i>Server-Based</i>	53
4.2.2	Layanan <i>Serverless</i>	54
4.2.2.1	Kubernetes Engine.....	54
4.2.2.2	App Engine	56
4.2.2.3	Cloud Run.....	57
4.2.2.4	Cloud Functions.....	59
4.2.2.5	Vertex AI	60
4.2.3	Komparasi Performa <i>Memory Utilization</i>	62
4.3	<i>Latency</i>	63
4.3.1	Layanan <i>Server-Based</i>	63
4.3.2	Layanan <i>Serverless</i>	64
4.3.2.1	Kubernetes Engine.....	64
4.3.2.2	App Engine	65
4.3.2.3	Cloud Run.....	66
4.3.2.4	Cloud Functions.....	67
4.3.2.5	Vertex AI	68
4.3.3	Komparasi Performa <i>Latency</i>	70
4.4	<i>Pricing</i>	71
4.4.1	Layanan <i>Server-Based</i>	71
4.4.2	Layanan <i>Serverless</i>	72
4.4.2.1	Kubernetes Engine.....	72
4.4.2.2	App Engine	73
4.4.2.3	Cloud Run.....	74
4.4.2.4	Cloud Functions.....	74
4.4.2.5	Vertex AI	75
4.4.3	Komparasi <i>Pricing</i>	76
4.5	<i>Developer Experiences</i>	78
BAB V SIMPULAN, IMPLIKASI, DAN REKOMENDASI		82
5.1	Simpulan.....	82
5.2	Implikasi	82
5.3	Rekomendasi	83

DAFTAR PUSTAKA	84
LAMPIRAN.....	92
Lampiran 1. <i>Deployment Model Machine Learning</i>	92
Lampiran 2. Hasil Pengujian Performa Setiap Layanan	113
Lampiran 3. Administrasi Penelitian	141

DAFTAR GAMBAR

Gambar 2.1 Arsitektur Kubernetes (Kubernetes, 2023).....	12
Gambar 2.2 Struktur <i>Decision Tree</i> (Charbuty & Abdulazeez, 2021)	18
Gambar 2.3 Tampilan GUI Apache JMeter	21
Gambar 2.4 Dashboard Pengujian JMeter Mode CLI	22
Gambar 2.5 Tampilan <i>dashboard</i> Cloud Monitoring	23
Gambar 3.1 Diagram Alur Penelitian	30
Gambar 3.2 Desain Arsitektur <i>Deployment</i> Compute Engine.....	32
Gambar 3.3 Desain Arsitektur <i>Deployment</i> App Engine	33
Gambar 3.4 Desain Arsitektur <i>Deployment</i> Kubernetes Engine	33
Gambar 3.5 Desain Arsitektur <i>Deployment</i> Cloud Run	34
Gambar 3.6 Desain Arsitektur <i>Deployment</i> Cloud Functions	34
Gambar 3.7 Desain Arsitektur <i>Deployment</i> Vertex AI	35
Gambar 4.1 <i>Test Plan</i> JMeter	41
Gambar 4.2 CPU <i>Utilization</i> Layanan Compute Engine.....	42
Gambar 4.3 Log Compute Engine	43
Gambar 4.4 CPU <i>Utilization</i> Layanan Kubernetes Engine	45
Gambar 4.5 CPU <i>Utilization</i> Layanan App Engine	46
Gambar 4.6 CPU <i>Utilization</i> Layanan Cloud Run	48
Gambar 4.7 CPU <i>Utilization</i> Layanan Cloud Functions	49
Gambar 4.8 CPU <i>Utilization</i> Layanan Vertex AI.....	51
Gambar 4.9 Komparasi Performa CPU <i>Utilization</i>	52
Gambar 4.10 <i>Memory Utilization</i> Layanan Compute Engine	54
Gambar 4.11 <i>Memory Utilization</i> Layanan Kubernetes Engine	55
Gambar 4.12 <i>Memory Utilization</i> Layanan App Engine.....	57
Gambar 4.13 <i>Memory Utilization</i> layanan Cloud Run.....	58
Gambar 4.14 <i>Memory Utilization</i> Layanan Cloud Functions	60
Gambar 4.15 <i>Memory Utilization</i> Layanan Vertex AI.....	61
Gambar 4.16 Komparasi <i>Memory Utilization</i>	62
Gambar 4.17 <i>Latency</i> Layanan Compute Engine.....	64
Gambar 4.18 <i>Latency</i> Layanan Kubernetes Engine	65
Gambar 4.19 <i>Latency</i> Layanan App Engine.....	66

Gambar 4.20 <i>Latency</i> Layanan Cloud Run	67
Gambar 4.21 <i>Latency</i> Layanan Cloud Functions	68
Gambar 4.22 <i>Latency</i> Layanan Vertex AI.....	69
Gambar 4.23 Komparasi <i>Latency</i>	71
Gambar 4.24 Komparasi <i>Pricing</i> Setiap Layanan	76

DAFTAR TABEL

Tabel 2.1 <i>Pricing</i> Layanan Compute Engine	10
Tabel 2.2 <i>Pricing</i> Layanan Kubernetes Engine.....	12
Tabel 2.3 <i>Pricing</i> Layanan App Engine	14
Tabel 2.4 <i>Pricing</i> Layanan Cloud Run.....	14
Tabel 2.5 <i>Pricing</i> Layanan Cloud Functions.....	15
Tabel 2.6 <i>Pricing</i> Layanan Vertex AI.....	16
Tabel 2.7 Standardisasi <i>Latency</i> Berdasarkan TIPHON	19
Tabel 2.8 Penelitian Relevan.....	23
Tabel 3.1 Konfigurasi Sistem Setiap Layanan	35
Tabel 4.1 Hasil Pengukuran CPU <i>Utilization</i> Layanan Compute Engine.....	42
Tabel 4.2 Hasil Pengukuran CPU <i>Utilization</i> Layanan Kubernetes Engine	44
Tabel 4.3 Hasil Pengukuran CPU <i>Utilization</i> Layanan App Engine	46
Tabel 4.4 Hasil Pengukuran CPU <i>Utilization</i> Layanan Cloud Run	47
Tabel 4.5 Hasil Pengukuran CPU <i>Utilization</i> Layanan Cloud Functions	48
Tabel 4.6 Hasil Pengukuran CPU <i>Utilization</i> Layanan Vertex AI.....	50
Tabel 4.7 Komparasi Performa CPU <i>Utilization</i>	51
Tabel 4.8 Hasil Pengukuran <i>Memory Utilization</i> Layanan Compute Engine	53
Tabel 4.9 Hasil Pengukuran <i>Memory Utilization</i> Layanan Kubernetes Engine... ..	54
Tabel 4.10 Hasil Pengukuran <i>Memory Utilization</i> Layanan App Engine	56
Tabel 4.11 Hasil Pengukuran <i>Memory Utilization</i> Layanan Cloud Run.....	57
Tabel 4.12 Hasil Pengukuran <i>Memory Utilization</i> Layanan Cloud Functions.....	59
Tabel 4.13 Hasil Pengukuran <i>Memory Utilization</i> Layanan Vertex AI	60
Tabel 4.14 Komparasi <i>Memory Utilization</i>	62
Tabel 4.15 Hasil Pengukuran <i>Latency</i> Layanan Compute Engine	63
Tabel 4.16 Hasil Pengukuran <i>Latency</i> Layanan Kubernetes Engine	64
Tabel 4.17 Hasil Pengukuran <i>Latency</i> Layanan App Engine.....	65
Tabel 4.18 Hasil Pengukuran <i>Latency</i> Layanan Cloud Run	66
Tabel 4.19 Hasil Pengukuran <i>Latency</i> Layanan Cloud Functions	68
Tabel 4.20 Hasil Pengukuran <i>Latency</i> Layanan Vertex AI.....	69
Tabel 4.21 Komparasi <i>Latency</i>	70
Tabel 4.22 <i>Pricing</i> Penggunaan Layanan Compute Engine.....	72

Tabel 4.23 <i>Pricing</i> Penggunaan Layanan Kubernetes Engine	73
Tabel 4.24 <i>Pricing</i> Penggunaan Layanan App Engine	73
Tabel 4.25 <i>Pricing</i> Penggunaan Layanan Cloud Run	74
Tabel 4.26 <i>Pricing</i> Penggunaan Layanan Cloud Functions	75
Tabel 4.27 <i>Pricing</i> Penggunaan Layanan Vertex AI	75
Tabel 4.28 Komparasi <i>Pricing</i> Penggunaan Setiap Layanan	76
Tabel 4.29 Komparasi <i>Price to Performance</i>	77
Tabel 4.30 Komparasi Performa Semua Layanan	80
Tabel 4.31 Ringkasan Hasil Komparasi Performa Layanan	80

DAFTAR LAMPIRAN

Lampiran 1.1. <i>Deployment</i> Menggunakan Compute Engine	92
Lampiran 1.2. <i>Deployment</i> Menggunakan Kubernetes Engine.....	96
Lampiran 1.3. <i>Deployment</i> Menggunakan App Engine	100
Lampiran 1.4. <i>Deployment</i> Menggunakan Cloud Run.....	102
Lampiran 1.5. <i>Deployment</i> Menggunakan Cloud Functions.....	104
Lampiran 1.6. <i>Deployment</i> Menggunakan Vertex AI	106
Lampiran 2.1. <i>CPU Utilization</i>	114
Lampiran 2.2. <i>Memory Utilization</i>	120
Lampiran 2.3. <i>Throughput</i>	126
Lampiran 2.4. <i>Latency</i>	138
Lampiran 3.1. Kartu Bimbingan Pembimbing 1	141
Lampiran 3.2. Kartu Bimbingan Pembimbing 2	142

DAFTAR PUSTAKA

- Abraham, A., & Yang, J. (2023a). A Comparative Analysis of Performance and Usability on Serverless and Server-Based Google Cloud Services. Dalam K. Daimi & A. Al Sadoon (Ed.), *Proceedings of the 2023 International Conference on Advances in Computing Research (ACR'23)* (Vol. 700, hlm. 408–422). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33743-7_33
- Abraham, A., & Yang, J. (2023b). *Analyzing the System Features, Usability, and Performance of a Containerized Application on Serverless Cloud Computing Systems*. Research Square. <https://doi.org/10.21203/rs.3.rs-3167840/v1>
- Al-Maliki, S. Q. A.-K. (2023). Efficient Cloud-Based Resource Sharing Through Multi-Tenancy and Load Balancing: An Exploration of Higher Education and Digital Libraries. *Iraqi Journal of Science*, 64(8), 4147–4159. <https://doi.org/10.24996/ij.s.2023.64.8.35>
- Altamevia, F., & Wijaya, H. O. L. (2023). Analisis Pola Penjualan Obat di Apotek Srikandi Menggunakan Algoritma Supervised Learning. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, 4(1), 170–176.
- Apache Software Foundation. (2024). *Apache JMeter—Apache JMeter™*. <https://jmeter.apache.org/>
- Arbain, A., Muhammad, M. A., Septiana, T., & Septama, H. D. (2022). Komparasi Implementasi Model Machine Learning Hoax News pada Local dan Cloud Computing Deployment Menggunakan Google App Engine. *Jurnal Informatika dan Teknik Elektro Terapan*, 10(3), 183–192. <https://doi.org/10.23960/jitet.v10i3.2646>
- Arifwidodo, B., Metayasha, V., & Ikhwan, S. (2021). Analisis Kinerja Load Balancing pada Server Web Menggunakan Algoritma Weighted Round Robin pada Proxmox VE. *Jurnal Telekomunikasi dan Komputer*, 11(3), 210. <https://doi.org/10.22441/incomtech.v11i3.11775>

- Azzahro, I. A., & Sofro, A. (2023). Regresi Robust untuk Pemodelan Deforestasi di Indonesia. *MATHunesa: Jurnal Ilmiah Matematika*, 11(3), 496–507. <https://doi.org/10.26740/mathunesa.v11n3.p496-507>
- Barua, H. B. (2021). *Data science and Machine learning in the Clouds: A Perspective for the Future* (Distributed, Parallel, and Cluster Computing arXiv:2109.01661). arXiv. <https://doi.org/10.48550/arXiv.2109.01661>
- Casalicchio, E. (2019). A study on performance measures for auto-scaling CPU-intensive containerized applications. *Cluster Computing*, 22(3), 995–1006. <https://doi.org/10.1007/s10586-018-02890-1>
- Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28. <https://doi.org/10.38094/jastt20165>
- Christidis, A., Davies, R., & Moschogiannis, S. (2019). Serving Machine Learning Workloads in Resource Constrained Environments: A Serverless Deployment Example. *2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA)*, 55–63. <https://doi.org/10.1109/SOCA.2019.00016>
- Cole, B. S., & Moore, J. H. (2018). Eleven quick tips for architecting biomedical informatics workflows with cloud computing. *PLOS Computational Biology*, 14(3), e1005994. <https://doi.org/10.1371/journal.pcbi.1005994>
- Dang, T. K., Mohan, N., Corneo, L., Zavodovski, A., Ott, J., & Kangasharju, J. (2021). Cloudy with a chance of short RTTs: Analyzing cloud connectivity in the internet. *Proceedings of the 21st ACM Internet Measurement Conference*, 62–79. <https://doi.org/10.1145/3487552.3487854>
- de Moraes, L. B., Parpinelli, R. S., & Fiorese, A. (2022). Application of deterministic, stochastic, and hybrid methods for cloud provider selection. *Journal of Cloud Computing*, 11(1), 5. <https://doi.org/10.1186/s13677-021-00275-1>
- Dianti, P. P. (2023). *Analisis Perbandingan Performa Container Pada Kubernetes Service Di Platform Google Cloud Platform (GCP), Amazon Web Service (AWS), Dan Microsoft Azure* [Skripsi]. Universitas Pendidikan Indonesia.

- Eisa, M., Younas, M., Basu, K., & Awan, I. (2020). Modelling and Simulation of QoS-Aware Service Selection in Cloud Computing. *Journal of Simulation Modelling Practice and Theory*, 103, 102108. <https://doi.org/10.1016/j.simpat.2020.102108>
- Erinle, B. (2017). *Performance Testing with JMeter 3* (Third). Packt Publishing.
- Fadil, A. (2020). Strategi Efisiensi Energi dan Penyeimbangan Beban Kerja Layanan Cloud Computing Melalui Konsolidasi Mesin Virtual Dinamis. *Applied Technology and Computing Science Journal*, 3(1), 1–12. <https://doi.org/10.33086/atcsj.v3i1.1680>
- Google Cloud. (2023a). *App Engine Application Platform | Google Cloud* [Documentation]. App Engine Application Platform | Google Cloud. <https://cloud.google.com/appengine?hl=en>
- Google Cloud. (2023b). *Cloud Functions | Google Cloud* [Documentation]. Cloud Functions | Google Cloud. <https://cloud.google.com/functions?hl=en>
- Google Cloud. (2023c). *Cloud Run* [Documentation]. Cloud Run. <https://cloud.google.com/run?hl=id>
- Google Cloud. (2023d). *Compute Engine | Google Cloud* [Documentation]. Compute Engine | Google Cloud. <https://cloud.google.com/compute?hl=en>
- Google Cloud. (2023e). *Google Kubernetes Engine (GKE)* [Documentation]. Google Kubernetes Engine (GKE). <https://cloud.google.com/kubernetes-engine?hl=en>
- Google Cloud. (2023f). *Lokasi Global—Region & Zona | Google Cloud* [Documentation]. Global Locations - Regions & Zones | Google Cloud. <https://cloud.google.com/about/locations?hl=id>
- Google Cloud. (2023g). *Vertex AI* [Documentation]. Vertex AI. <https://cloud.google.com/vertex-ai?hl=en>
- Google Cloud. (2024a). *About VM metadata* [Documentation]. <https://cloud.google.com/compute/docs/metadata/overview>
- Google Cloud. (2024b). *All networking pricing | Virtual Private Cloud* [Pricing]. Google Cloud. <https://cloud.google.com/vpc/network-pricing>

- Google Cloud. (2024c). *Arsitektur cluster GKE* [Documentation]. Google Cloud. <https://cloud.google.com/kubernetes-engine/docs/concepts/cluster-architecture?hl=id>
- Google Cloud. (2024d). *Artifact Registry Pricing* [Pricing]. Google Cloud. <https://cloud.google.com/artifact-registry/pricing>
- Google Cloud. (2024e). *Cloud Functions Pricing* [Pricing]. Google Cloud. <https://cloud.google.com/functions/pricing>
- Google Cloud. (2024f). *Cloud Monitoring*. <https://cloud.google.com/monitoring?hl=en>
- Google Cloud. (2024g). *Cloud Run Pricing* [Pricing]. <https://cloud.google.com/run/pricing>
- Google Cloud. (2024h). *Cloud Storage Pricing* [Pricing]. Google Cloud. <https://cloud.google.com/storage/pricing>
- Google Cloud. (2024i). *Compute Engine: Virtual Machines (VMs) Pricing* [Pricing]. <https://cloud.google.com/compute/all-pricing#compute-engine-pricing>
- Google Cloud. (2024j). *Google Kubernetes Engine (GKE) Pricing* [Pricing]. Google Cloud. <https://cloud.google.com/kubernetes-engine/pricing>
- Google Cloud. (2024k). *Guest environment* [Documentation]. Google Cloud. <https://cloud.google.com/compute/docs/images/guest-environment>
- Google Cloud. (2024l). *Mengoptimalkan latensi aplikasi dengan load balancing* [Documentation]. Google Cloud. <https://cloud.google.com/load-balancing/docs/tutorials/optimize-app-latency?hl=id>
- Google Cloud. (2024m). *Metrics, time series, and resources* [Documentation]. Google Cloud. <https://cloud.google.com/monitoring/api/v3/metrics>
- Google Cloud. (2024n). *Retention and latency of metric data* [Documentation]. Google Cloud. <https://cloud.google.com/monitoring/api/v3/latency-n-retention>
- Google Cloud. (2024o). *Storage options* [Documentation]. Storage options. <https://cloud.google.com/compute/docs/disks>
- Google Cloud. (2024p). *Vertex AI Pricing* [Pricing]. https://cloud.google.com/vertex-ai/pricing#automl_models

- Google Cloud. (2024q). *View Google Cloud latency dashboard | Performance Dashboard* [Documentation]. Google Cloud. <https://cloud.google.com/network-intelligence-center/docs/performance-dashboard/how-to/view-google-cloud-latency>
- Google Cloud. (2024r). *VM Manager | Dokumentasi Compute Engine* [Pricing]. Google Cloud. <https://cloud.google.com/compute/docs/vm-manager?hl=id>
- Hao, T., Zhan, J., Hwang, K., Gao, W., & Wen, X. (2021). AI-oriented Workload Allocation for Cloud-Edge Computing. *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 555–564. <https://doi.org/10.1109/CCGrid51090.2021.00065>
- Hawkins, D. M. (1980). Introduction. Dalam *Identification of Outliers* (1 ed., hlm. 1). Springer Dordrecht.
- Heryana, N. (2018). Analisis Pemanfaatan Cloud Computing Berbasis Software as a Service sebagai Media Penyimpanan Tugas Praktikum. *SYNTAX Jurnal Informatika*, 7(1), 26–32.
- IBM. (2022a). *Cost per Transaction—Informix Servers 14.10* [Documentation]. <https://www.ibm.com/docs/en/informix-servers/14.10?topic=performance-cost-per-transaction>
- IBM. (2022b). *CPU utilization—Informix Servers 14.10* [Documentation]. <https://www.ibm.com/docs/en/informix-servers/14.10?topic=performance-cost-per-transaction>
- IBM. (2022c). *Memory Utilization—Informix Servers 14.10* [Documentation]. <https://www.ibm.com/docs/en/informix-servers/14.10?topic=performance-memory-utilization>
- Ibrahim, A., Alang, A. H., Madi, Baharuddin, Ahmad, M. A., & Darmawati. (2018). *Metodologi Penelitian*. Gunadarma Ilmu.
- Ifrah, S. (2021). *Getting Started with Containers in Google Cloud Platform*. Apress.
- Jang, H.-C., & Luo, S.-Y. (2023). Enhancing Node Fault Tolerance through High-Availability Clusters in Kubernetes. *2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, 30–35. <https://doi.org/10.1109/ICEIB57887.2023.10170110>

- Jeong, B., Jeon, J., & Jeong, Y.-S. (2023). Proactive Resource Autoscaling Scheme Based on SCINet for High-Performance Cloud Computing. *IEEE Transactions on Cloud Computing*, *11*(4), 3497–3509. IEEE Transactions on Cloud Computing. <https://doi.org/10.1109/TCC.2023.3292378>
- Kubernetes. (2023). *Cluster Architecture*. <https://kubernetes.io/docs/concepts/architecture/>
- Kubernetes. (2024). *Kubernetes Documentation*. <https://kubernetes.io/docs/home/>
- Kurniawan, D. (2022). *Pengenalan Machine Learning dengan Python*. PT Elex Media Komputindo.
- Lakkadwala, P., & Kanungo, P. (2018). Memory Utilization Techniques for Cloud Resource Management in Cloud Computing Environment: A Survey. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–5. <https://doi.org/10.1109/CCAA.2018.8777457>
- Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157–170. <https://doi.org/10.1016/j.bushor.2019.10.005>
- Mancini, R., Cucinotta, T., & Abeni, L. (2020). *Performance Modeling in Predictable Cloud Computing*. 69–78. <https://doi.org/10.5220/0009349400690078>
- Marinescu, D. C. (2018). *Cloud Computing Theory and Practice* (Second Edition). Morgan Kaufmann.
- Mason, K., Duggan, M., Barrett, E., Duggan, J., & Howley, E. (2018). Predicting host CPU utilization in the cloud using evolutionary neural networks. *Future Generation Computer Systems*, *86*, 162–173. <https://doi.org/10.1016/j.future.2018.03.040>
- Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing*. NIST.
- Miyachi, C. (2018). What is “Cloud”? It is time to update the NIST definition? *IEEE Cloud Computing*, *5*(3), 6–11. <https://doi.org/10.1109/MCC.2018.032591611>
- Mohammed, C. M., & Zeebaree, S. R. M. (2021). Sufficient Comparison Among Cloud Computing Services: IaaS, PaaS, and SaaS: A Review. *International*

- Journal of Science and Business*, 5(2), 17–30.
<https://doi.org/10.5281/ZENODO.4450129>
- Mohan, V. M., Yaraguti, R., Chinta, S. S., & Jonnavithula, B. (2024). Optimizing Cloud Load Balancers for Reduced Network Latency. Dalam *Algorithms in Advanced Artificial Intelligence* (1 ed., hlm. 6). CRC Press.
- Molyneaux, I. (2009). *The Art of Application Performance Testing* (First). O'Reilly Media.
- Pratama, H. P., Putri, D. I. H., Putri, H. E., Irawan, E. N., & Kautsar, M. A. (2024). Smart watering of ornamental plants: Exploring the potential of decision trees in precision agriculture based on IoT. *Journal of Mechatronics, Electrical Power, and Vehicular Technology*, 15(1), 82–92.
- Rahman, F. (2023). *Serverless Cloud Computing: A Comparative Analysis of Performance, Cost, and Developer Experiences in Container-Level Services* [Thesis]. Aalto University School of Science.
- Ramsari, N., & Ginanjar, A. (2022). Implementasi Infrastruktur Server Berbasis Cloud Computing Untuk Web Service Berbasis Teknologi Google Cloud Platform. *Conference SENATIK STT Adisutjipto Yogyakarta*, 7, 169–182.
<https://doi.org/10.28989/senatik.v7i0.472>
- Riana, E. (2020). Implementasi Cloud Computing Technology dan Dampaknya Terhadap Kelangsungan Bisnis Perusahaan Dengan Menggunakan Metode Agile dan Studi Literatur. *JURIKOM (Jurnal Riset Komputer)*, 7(3), 439–449. <https://doi.org/10.30865/jurikom.v7i3.2192>
- Sahabuddin, R., Idrus, M. I., & Karim, A. (2021). Pengukuran Nilai Sentral dan Jenisnya. Dalam *Pengantar Statistika* (1 ed., hlm. 103–104). Liyan Pustaka Ide.
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 160.
<https://doi.org/10.1007/s42979-021-00592-x>
- Sihombing, P. R., Suryadiningrat, S., Sunarjo, D. A., & Yuda, Y. P. A. C. (2023). Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya. *Jurnal Ekonomi Dan Statistik Indonesia*, 2(3), 307–316. <https://doi.org/10.11594/jesi.02.03.07>

- Statista Research Department. (2023). *Machine learning*. Statista. <https://www-statista-com.i.ezproxy.nypl.org/study/111190/machine-learning/>
- Statista Research Department. (2024). *Indonesia: Government AI readiness index by category 2022*. Statista. <https://www-statista-com.i.ezproxy.nypl.org/statistics/1411083/indonesia-government-ai-readiness-index-by-category/>
- Sugiyono. (2013). *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Alfabeta.
- Syarafina, M. (2023). *Analisis Perbandingan Serverless Deployment Machine Learning Menggunakan App Engine dan Cloud Function pada Aplikasi Spezia* [Skripsi]. Universitas Jenderal Soedirman.
- Wu, Y., Dinh, T. T. A., Hu, G., Zhang, M., Chee, Y. M., & Ooi, B. C. (2022). Serverless Data Science - Are We There Yet? A Case Study of Model Serving. *Proceedings of the 2022 International Conference on Management of Data*, 1866–1875. <https://doi.org/10.1145/3514221.3517905>
- Xu, R. (2020). *A Design Pattern for Deploying Machine Learning Models to Production*. Computer Science and Information Systems California State University San Marcus. <https://scholarworks.calstate.edu/downloads/1v53k296v>
- Zhang, Z., Yang, C., Wang, J., Hou, G., & Yang, Y. (2022). Scalable Resource Provisioning For Multi-tenant SaaS With Cloud Functions. *2022 International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*, 181–187. <https://doi.org/10.1109/CCPQT56151.2022.00039>
- Zhu, W., Hou, A. B., Yang, R., Datta, A., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). QuakeFlow: A scalable machine-learning-based earthquake monitoring workflow with cloud computing. *Geophysical Journal International*, 232(1), 684–693. <https://doi.org/10.1093/gji/ggac355>