INTISARI

Platform as a Service (PaaS) dalam praktik pengelolaan aplikasi masih menghadapi tantangan pada penentuan kapasitas atau ukuran layanan (*cloud sizing*), yang sering dilakukan secara reaktif atau berbasis evaluasi metrik sesaat. Pendekatan tersebut menyebabkan keputusan kapasitas mudah berubah ketika terjadi fluktuasi beban jangka pendek, sehingga meningkatkan risiko *under-provisioning* dan *over-provisioning.* Kajian terdahulu telah berupaya mengatasi risiko tersebut melalui *workload forecasting*, *threshold-based autoscaling*, dan *risk-aware provisioning*, namun integrasi pendekatan tersebut ke dalam keputusan kapasitas PaaS pra-*runtime* secara *end-to-end* masih terbatas. Penelitian ini bertujuan merancang dan mengimplementasikan sistem *Cloud Sizing Automation* dengan metode R²D (*Risk-Aware Rule-Based Decision*) untuk menghasilkan keputusan kapasitas yang lebih adaptif, stabil, terukur, dan dapat diaudit pada tahap *pra-runtime.*

Penelitian ini menggunakan dua *dataset* agregat utama metrik historis App Service dan Azure SQL, yaitu App Service Juni-Agustus 2025 sebanyak 117.984 baris dan SQL Juni-Agustus 2025 sebanyak 118.920 baris, sehingga total data historis utama yang digunakan dalam evaluasi inti adalah 236.904 baris pada periode Juni-Agustus 2025. *Pipeline* sistem mencakup validasi metadata dan *signal quality gate*, pemodelan *forecasting* menggunakan AR, ARIMA, SARIMA, RNN, GRU, dan LSTM dengan *baseline* Seasonal Naive, evaluasi akurasi menggunakan RMSE, MAE, dan R^2, serta pengambilan keputusan berbasis R²D melalui analisis *temporal, risk scoring over/under-provisioning, capacity ladder,* dan *guardrail* kompatibilitas. Output keputusan diintegrasikan dengan *webhook, runbook, scheduler*, dan *closed-loop runtime callback* agar keputusan dapat ditelusuri hingga status eksekusi.

Hasil pengujian menunjukkan bahwa *quality gate* meloloskan 2 dari 2 dataset agregat utama. Pada scope App Service, model terbaik menghasilkan kinerja CPU sebesar RMSE 12,0024, MAE 6,7756, dan R^2 0,6386, sedangkan pada sinyal memory menghasilkan RMSE 2,4694, MAE 1,3290, dan R^2 0,9651. Pada scope Azure SQL, model terbaik menghasilkan RMSE 3,2105, MAE 1,3398, dan R^2 0,2216 pada sinyal DTU. Pada set evaluasi inti, sistem menghasilkan 32 keputusan dengan distribusi aksi seluruhnya *DOWN* dan *confidence* rata-rata 90,18%. Seluruh keputusan pada set inti berstatus *executed* serta tersinkronisasi melalui *callback*, sehingga *execution rate* dan *automation success rate* pada set ini mencapai 100%. Temuan ini menunjukkan bahwa R²D efektif mentransformasikan hasil *forecasting* menjadi keputusan kapasitas yang eksplisit, konsisten, dan dapat ditelusuri secara *end-to-end* pada konteks layanan PaaS.

Kata kunci—*Platform as a Service (PaaS), Cloud Sizing, Forecasting, Risk-Aware Rule-Based Decision, Capacity Management, Closed-Loop Automation.*

ABSTRACT

In practice, Platform as a Service (PaaS) still faces significant challenges in determining appropriate service capacity or size (cloud sizing) for application management, which is often handled reactively or based only on instantaneous metric evaluations. This approach makes capacity decisions highly sensitive to short-term workload fluctuations, increasing the risks of under-provisioning and over-provisioning. Previous studies have attempted to address these risks through workload forecasting, threshold-based autoscaling, and risk-aware provisioning; however, the integration of these approaches into end-to-end pre-runtime PaaS capacity decisions remains limited. This study aims to design and implement a Cloud Sizing Automation system using the R²D (Risk-Aware Rule-Based Decision) method to produce capacity decisions that are more adaptive, stable, measurable, and auditable at the pre-runtime stage.

The research uses two main aggregated historical datasets from App Service metrics and Azure SQL metrics, namely App Service June-August 2025 with 117,984 records and SQL June-August 2025 with 118,920 records, resulting in a total of 236,904 historical records used in the core evaluation during June-August 2025. The system pipeline includes metadata validation and a signal quality gate, forecasting models using AR, ARIMA, SARIMA, RNN, GRU, and LSTM with Seasonal Naive as the baseline, accuracy evaluation using RMSE, MAE, and R², and R²D-based decision-making through temporal analysis, over/under-provisioning risk scoring, capacity ladder selection, and compatibility guardrails. Decision outputs are integrated with webhook, runbook, scheduler, and closed-loop runtime callbacks to ensure traceability through execution status.

The testing results show that the quality gate accepts two out of two main aggregated datasets. In the App Service scope, the best model achieves CPU forecasting performance of RMSE 12.0024, MAE 6.7756, and R² 0.6386, while memory forecasting achieves RMSE 2.4694, MAE 1.3290, and R² 0.9651. In the Azure SQL scope, the best model achieves RMSE 3.2105, MAE 1.3398, and R² 0.2216 on the DTU signal. In the core evaluation set, the system produced 32 decisions, all classified as DOWN, with an average confidence of 90.18%. All decisions in this set reached executed status and were synchronized via callback, resulting in an execution rate and automation success rate of 100.00% for this set. These findings indicate that R²D effectively transforms forecasting outputs into explicit, consistent, and end-to-end traceable capacity decisions in a PaaS context.

Keywords*—Platform as a Service (PaaS), Cloud Sizing, Forecasting, Risk-Aware Rule-Based Decision, Capacity Management, Closed-Loop Automation.*