

Raccolta e integrazione di telemetria applicativa per sistemi distribuiti Cloud-Edge

Proposta

Le decisioni di orchestrazione nei sistemi distribuiti non dipendono solo dalle caratteristiche della rete, ma anche dal comportamento delle applicazioni (volume delle richieste tra servizi, utilizzo di CPU/memoria, latenze di risposta, variabilità del carico e pattern temporali, come picchi, ciclicità giornaliere o comportamenti bursty).

Questi fattori influenzano direttamente scelte come il posizionamento dei servizi, il numero di repliche o la necessità di migrazioni tra nodi cloud ed edge.

Nello stato dell'arte esistono tracce reali di workload e telemetria applicativa provenienti da cluster e benchmark di microservizi, ma tali dati sono spesso isolati, raccolti con metriche diverse, formati eterogenei e con livelli di dettaglio non uniformi. Questo rende difficile riutilizzarli in modo sistematico per studi su sistemi distribuiti, orchestrazione e gestione delle risorse.

L'obiettivo della tesi è raccogliere, analizzare e integrare telemetria applicativa reale costruendo un dataset coerente, pulito e strutturato per lo studio di sistemi distribuiti cloud-edge. Il dataset risultante dovrà rappresentare in modo omogeneo servizi, istanze applicative, richieste, metriche di utilizzo e tempi di risposta, così da poter essere utilizzato come base per simulazioni, studi di gestione delle risorse e valutazione di tecniche di orchestrazione.

Il lavoro prevede l'analisi di benchmark e tracce pubblici di applicazioni distribuite, la definizione di un modello dati unificato per descrivere servizi e metriche, la pulizia dei dati (gestione di valori mancanti, unità di misura, formati) e l'allineamento temporale tra diverse sorgenti. Una parte del Lavoro di tesi sarà dedicato alla caratterizzazione dei pattern applicativi, ad esempio studiando la variabilità del carico nel tempo, la relazione tra utilizzo delle risorse e latenza, e la presenza di picchi o comportamenti anomali. Il risultato sarà un dataset documentato e riutilizzabile, accompagnato da un'analisi che ne descriva le principali proprietà.

Strumenti e tecnologie

- Python
- Pandas / NumPy
- Benchmark di microservizi e tracce pubblici
- Matplotlib / Seaborn
- JSON / CSV / Parquet

Obiettivi

1. Analizzare fonti dati di telemetria applicativa.
2. Progettare uno schema dati unificato per servizi e metriche.
3. Implementare pipeline di integrazione e pulizia.
4. Studiare pattern di carico e utilizzo risorse.
5. Produrre un dataset documentato per sistemi distribuiti.
6. È inoltre richiesto di mostrare un caso d'uso concreto che utilizzi il dataset prodotto.

Prerequisiti

- Buona conoscenza di Python.
- Basi di sistemi operativi e sistemi distribuiti.
- Interesse per analisi dati e cloud-edge computing.