Petar Prlina

Sistem za predlaganje cloud servera i infrastrukture

Dolazak na ideju

Odabir adekvatnog cloud servera je težak jer korisnici često ne poznaju dovoljno hardver, mrežu i modele naplate. Naša aplikacija omogućava i korisnicima bez stručnog znanja da izaberu optimalnu instancu za svoju upotrebu (web, baze, analitika, ML, streaming), a zatim da je jednostavno zakupe.

Problem

Većina cloud provajdera prikazuje katalog instance i prepušta korisniku da sam proceni šta mu treba, što često dovodi do preplaćivanja ili slabih performansi. Naš servis koristi znanjem vođenu pretragu (Drools pravila) kako bi, na osnovu potreba i budžeta, rangirao i predložio najadekvatnije konfiguracije, preko više provajdera i regiona.

Uloge u sistemu

Korisnik

- Registruje nalog i prijavljuje se da bi dobio personalizovane preporuke i istoriju zakupa.
- Popunjava interaktivnu formu sa poslovnim i tehničkim zahtevima (svrha korišćenja, budžet, traženi nivo zaštite, potreba za grafikom, ekološki prioritet) i dobija listu preporučenih instanci.
- Pokreće zahtev za zakup iz kataloga ili direktno sa preporuka; zahtev dobija oznaku „na čekanju” dok ga administrator ne obradi
- Na korisničkoj tabli prati sve zahteve i aktivne ugovore uz vidljive statuse „na čekanju”, „aktivno”, „odbijeno” ili „završeno”.
- Kada je instanca aktivna određeni period, može da unese ocenu i komentar
- Sa svakim uspešnim zakupom gradi lojalti nivo (bronza, srebro, zlato) koji zatim donosi popuste i dodatne bodove prilikom novih preporuka

Admin

- Pregleda listu zahteva koji čekaju odluku i odobrava ili odbija svaki pojedinačno.
- Uređuje katalog usluga: dodaje nove konfiguracije, menja cene, ažurira ponudu grafičkih kartica, skladišta i regiona.
- Prima obaveštenja u realnom vremenu kada analitika otkrije da je neka instanca previše loše ili izuzetno dobro ocenjena, pa može da ukloni ponudu ili zatraži dodatne kapacitete od provajdera.

Metodologija rada

Očekivani ulaz

- Svrha (web aplikacija, baza podataka, analitika podataka, ML trening/inferencija, streaming).
- CPU zahtev (niske/srednje/visoke performanse; minimalan broj vCPU).
- GPU zahtev (nema, 1+, minimalan VRAM).
- Memorija (minimalni GB).
- Skladište (kapacitet; tip: NVMe/SATA, šifrovanje).
- Mreža (minimalni protok/širina opsega, DDoS zaštita).
- Visoka dostupnost (da/ne; multi-zona).
- Region i usaglašenost (EU/US/APAC; GDPR, ISO 27001).
- Eko prioritet (da/ne).
- Očekivani broj istovremenih korisnika/opterećenje (npr. <200, 200-1000, 1000+).
- Budžet (niski, srednji, visok) i model naplate (po satu, mesečno).
- Trajanje zakupa (broj_dana).

Očekivani izlaz

- Konfiguracije koje najbolje zadovoljavaju unose, rangirane po bodovima, sa procenjenim mesečnim troškom

Pravila

Pravila pretrage

- Svaka instanca kreće sa početnim brojem bodova, a zatim prolazi kroz seriju pravila koja proveravaju procesorske, grafičke, memorijske, skladišne, mrežne i sigurnosne zahteve. Ispunjavanje uslova donosi bodove, a propuštanje ih umanjuje.
- Trajanje zakupa donosi progresivne popuste: preko mesec dana uračunava 5 %, preko tri meseca 10 %, a preko šest meseci 15 %. Ovi popusti se spajaju sa lojalti nivoima kroz proračun koji sprečava dvostruko računanje.
- Specijalizovani scenariji (intenzivni ML trening, kritične baze, mali web saobraćaj, ekološki prioritet striminga) uvode dodatne oznake koje otključavaju zasebna pravila i nagrađuju konfiguracije sa odgovarajućim karakteristikama poput NVMe skladišta, višezonske redundanse ili energetske efikasnosti.
- Budžetska pravila proveravaju krajnju cenu i nagrađuju kombinacije koje ostaju u zadatim okvirima, dok se neadekvatne ponude eliminišu pre završnog rangiranja.

Pravila za dobijanje statusa povlastica

- Korisnici koji iznajme server najmanje 3 puta dobijaju status bronzani.
- Korisnici koji iznajme server najmanje 5 puta dobijaju status srebrni.
- Korisnici koji iznajme server najmanje 7 puta dobijaju status zlatni.

- Na osnovu statusa korisnici dobijaju popust pri zakupu od 5%, 10% ili 15% za svaki nivo.
- Korisnici koji iznajme server na više od mesec dana dobijaju popust od 5%.
- Korisnici koji iznajme server na više od 3 meseca dobijaju popust od 10%.
- Korisnici koji iznajme server na više od 6 meseci dobijaju popust od 15%.

Template za eliminaciju nepodobnih konfiguracija:

- Sistemska pravila za eliminaciju se generišu iz šablona koji obuhvata najčešće obavezne filtere. Šablon se učitava pri pokretanju aplikacije i pretvara u konkretna pravila bez potrebe za ručnim ažuriranjem DRL datoteka.
- Kada korisnik zada minimalne vrednosti za CPU, GPU, RAM, skladište, region ili budžet, odgovarajuća pravila iz šablona automatski uklanjaju sve ponude koje padnu ispod tog praga. Eliminišu se i konfiguracije bez zahtevane zaštite (DDoS, enkripcija) kako bi se kasnije bodovanje fokusiralo samo na realne kandidate.

Pravila za obaveštavanje admina (**complex event processing**)

- Analitički sloj za kompleksnu obradu događaja posmatra vremenski prozor od 30 dana u koji ulaze svi novi zakupi i ocene. Svaki novi događaj se analizira „u hodu”, bez čekanja batch obrade.
- Za svaki server računa se ponderisani rezultat. Sistem posebno prati sledeće signale:
- trajanje aktivnog zakupa (duže korišćenje povećava pouzdanost rezultata i težinu upozorenja),
- lojalti status korisnika (bronza, srebro, zlato) koji se koristi kao multiplikator uticaja njihove ocene,
- tempo ponovljenih kupovina, tj. „velocity” faktor koji raste kada korisnici češće obnavljaju zakup iste konfiguracije,
- prosečna ocena i njen skoriji trend; duži ugovori i viši status korisnika pojačavaju uticaj ovih ocena na krajnji signal.
- Kada rezultat padne ispod približno 20 procenata, sistem generiše upozorenje da performanse opadaju (npr. problemi sa kvalitetom usluge). Kada poraste iznad 70 procenata, kreira se pozitivan signal da instanca radi iznad očekivanja i da je možda vredno proširiti kapacitet.
- Obaveštenje sadrži naziv servera, provajdera i kontekstualnu poruku. Preko real-time kanala (notifikacije u admin panelu) administrator odmah dobija informaciju i može da reaguje

Ulančavanja pravila

- Ako je svrha ML trening i dataset je velik (npr. >100GB), generiše se činjenica “težak ML trening”.

        - Ako postoji činjenica “težak ML trening”, aktivira se pravilo koje dodaje bodove instancama sa barem jednim GPU-om sa visokim VRAM-om i NVMe skladištem većeg kapaciteta.
        - Ako postoji činjenica “težak ML trening”, aktivira se pravilo koje dodaje bodove instancama sa višim mrežnim propusnim opsegom i dediciranim CPU-om.
        - Ako postoji činjenica "težak ML trening" i budžet je "visok" i trajanje zakupa je 6+ meseci, generiše se činjenica "enterprise ML projekat".
            - Ako postoji činjenica "enterprise ML projekat" i region je EU sa GDPR zahtevom, aktivira se pravilo koje dodatno boduje provajdere sa on-premise opcijama i mogućnošću hybrid cloud deployment-a
- Ako je svrha web aplikacija i očekivano opterećenje je manje od 200 korisnika, generiše se činjenica “mali web saobraćaj”.
    - Ako postoji činjenica “mali web saobraćaj”, aktivira se pravilo koje dodaje bodove manjim instancama sa mogućnošću kasnijeg autoscalinga.
- Ako je svrha baza podataka i visoka dostupnost je označena, generiše se činjenica “kritična baza”.
    - Ako postoji činjenica “kritična baza”, aktivira se pravilo koje dodaje bodove menadžerisanim bazama sa replikacijom i NVMe diskovima sa visokim IOPS.
    - Ako postoji činjenica “kritična baza”, aktivira se pravilo koje preferira konfiguracije raspoređene preko više zona.
- Ako je svrha streaming i označen je eko prioritet, generiše se činjenica “eko streaming”.
    - Ako postoji činjenica “eko streaming”, aktivira se pravilo koje dodaje bodove konfiguracijama u zelenim data centrima i sa energetski efikasnim instancama.

Tehnologije

- Backend: Java Spring Boot + Drools
- Frontend: React