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

- Može da pretražuje sve servere po nazivu tipa, provajderu ili regionu.
- Može da pretražuje servere na osnovu svojih potreba (svrha, resursi, budžet).
- Mora da se registruje da bi mogao da rezerviše/provizionira server.
- Može da pregleda listu servera koje je zakupio.
- Može da oceni server/provajdera nakon korišćenja.
- Dobija povlastice na osnovu broja iznajmljivanja i/ili trajanja zakupa.

Admin

- Admin može da doda nove tipove servera, GPU modele, disk opcije, regione i provajdere.
- Admin odobrava ili odbija pristigle zahteve za zakup/provizioniranje.

Metodologija rada

Očekivani ulaz

- Svrha (web aplikacija, baza podataka, analitika podataka, ML trening/inferencija, streaming/transkodiranje).
- 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, automatski failover).
- Region i usaglašenost (EU/US/APAC; GDPR, ISO 27001).
- Eko prioritet (da/ne; prioritet zelenih/karbonski neutralnih data centara).

- Očekivani broj istovremenih korisnika/opterećenje (npr. <200, 200-1000, 1000+).
- Budžet (niski, srednji, visok) i model naplate (po satu, mesečno).
- Trajanje zakupa (broj_dana, 1+ mesec, 3+ meseca, 6+ meseci).

Očekivani izlaz

- Konfiguracije koje najbolje zadovoljavaju unose, rangirane po bodovima, sa procenjenim mesečnim troškom, dostupnim regionima i obrazloženjem (zašto je predlog dobar).

Pravila

Pravila pretrage

- Ako je označen eko prioritet:
    - Aktivira se pravilo koje dodaje bodove konfiguracijama u zelenim ili karbonski neutralnim data centrima.
- Unosi za dodatne zahteve (DdoS zaštita, šifrovanje, backup) aktiviraju pravila koja daju bodove instancama koje to podržavaju.
- Na osnovu svrhe biraju se adekvatne konfiguracije:
    - Web aplikacija: balans CPU/RAM; ako je HA tražena, prednost multi-zona rešenjima i managed load balancer-ima.
    - Baza podataka: više RAM-a, NVMe sa visokim IOPS; prednost menadžerisanim DB-ovima; ako je HA, zahtev za replikacijom.
    - Analitika: mnogo jezgara, veliki brzi disk (NVMe), mogućnost horizontalnog skaliranja.
    - ML trening: GPU obavezno (VRAM u skladu sa modelom), NVMe za dataset, veći bandwidth; inferencija daje bodove manjim GPU instancama sa niskom latencijom.
    - Streaming/transkodiranje: visok mrežni protok, opcionalno GPU za transkodiranje, stabilan IO.
- Personalizacija na osnovu istorije:
    - Ako korisnik, u proteklih godinu dana, ima najmanje 10 zakupa i u ≥20% slučajeva je birao jednog provajdera, dodavaće se bodovi svim konfiguracijama tog provajdera tokom pretrage.
    - Ako je korisnik, u proteklih godinu dana, ocenio najmanje 10 instanci/provajdera ocenom 4 ili 5, dodavaće se bodovi svim konfiguracijama tog provajdera tokom pretrage.

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:

- Omogućava automatsko generisanje pravila za eliminaciju konfiguracija koje ne zadovoljavaju zahteve. Template prima tip kriterijuma i način provere, te automatski uklanja neodgovarajuće ponude iz razmatranja.
- Primeri:
    - Na osnovu unosa za opterećenje (istovremeni korisnici):
        - Izbaciće se konfiguracije koje ne mogu da isporuče dovoljan throughput; dodaju se bodovi instancama sa adekvatnim CPU/mrežom.
    - Na osnovu unosa za region/usaglašenost:
        - Izbaciće se konfiguracije van izabranih regiona ili bez traženog sertifikata (npr. GDPR, ISO 27001).
    - Na osnovu budžeta:
        - Izbaciće se konfiguracije koje izlaze iz budžeta korisnika.

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

- Sistem kontinuirano analizira ocene i performanse različitih tipova instanci kroz više dimenzija - broj ocena i korelacije sa drugim faktorima i predviđa buduće potrebe.
- U odnosu na to se adminu generiše izveštaj svaki put kada jedan tip instance udje u ekstrem >80% ili <20% (gde je 100% sjajno a 0% loše), po sledećim kriterijumima, i preporučuje adminu ili da ukloni il da ih doda još.
    - Sve ocene, kao i ocene u poslednjih mesec dana, gledano ekvivalentno
    - Tip korisnika koji je dao ocenu (novi/postojeći, bronze/silver/gold status)
    - Trajanje korišćenja pre davanja ocene
    - Svrha korišćenja instance (web, DB, ML, itd.)
    - Trenutna iskorišćenost kapaciteta po tipu instance
    - Istorija promena kapaciteta

Primeri 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