Rangirajući pretraživač dokumenata - Infinity

Marko Budiselić

30.11.2015.

1 Predprocesiranje

2 Algoritmi

2.1 Bag of words

2.2 Vector space

2.3 Binary independence

3 Upute za pokretanje

`https://infinity.buda.link`

TODO: source setup.py

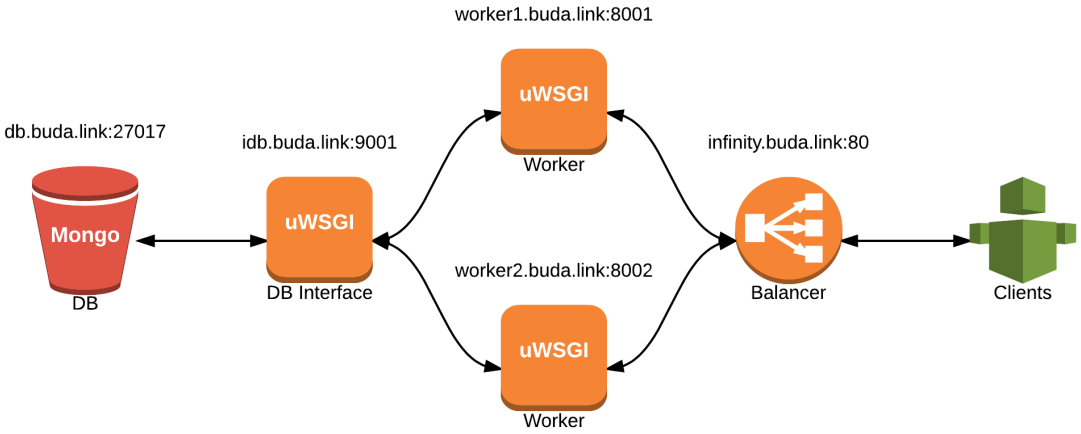
4 Produkcijska okolina

Svi upiti klijenata dolaze na Nginx load balancer koji ima izrazito veliku propusnost. Nakon toga load banancer prosljedjuje upit na worker instance koje svaka za sebe imaju cijeli dataset i znaju vratiti odgovarajuci rezultat. Dataset worker instance preuzimaju od db interface instance, a ne direktno iz baze. Trenutno je u deploymentu samo jedna instanca sucelja prema bazi, ali u produkciji tu moze biti opet load balancer i vise instanci interface-a prema bazi podataka. Baza podataka je MongoDB, takodjer samo jedna instanca, no u praksi tu moze doci mongo replica set ili mongo shard cluster.

Kao sto se vidi na grafu 1, sve instance imaju simbolicka imena (FQDN). To je takodjer izrazito bitno jer se time postize transparentnost pristupa i migracijska transparentnost. U konkretnoj implementaciji sva imena su definirana u /etc/hosts file-u na deploy stroju, no u produkciji ce imene biti definirana na redundantnim DNS serverima.

P.S.

infinity > gugol



Slika 1: Produkcijska okolina