RAG 서버 오라클 인스턴스 정보

Shape configuration

Shape: VM.Standard3.Flex

OCPU count: 8

Network bandwidth (Gbps): 8

Memory (GB): 16

Local disk: Block storage only

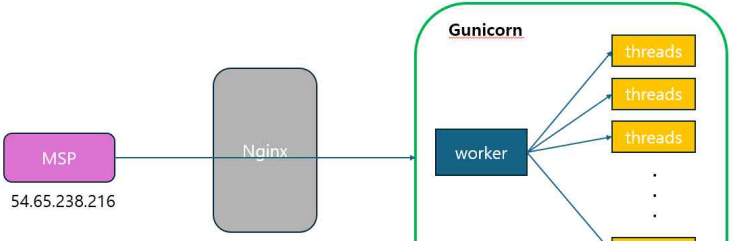
기본 RAG server 는 flask python server로 싱글스레드여서 동시 접속 진행시 fail이 나온 것을
기존에 pm2 설정에 Gunicorn을 추가함으로써 오라클 인스턴스 리소스 바탕으로
워커 8개, 워커 1개에 스레드 16개 설정해서 동시 접속 128개 허용할 수 있게 구현했습니다.

ocpu 기반으로 워커 및 스레드 설정 참조 문서 링크는 Gunicorn에서
https://docs.gunicorn.org/en/stable/settings.html#workers 참고해서 2 num_cores = 1 ocpu를 기반으로 워커
및 스레드 설정 8개, 스레드는 워커 1개 16개로 설정해서 구현 및 적용했습니다.

python locust 동시 접속 테스트를 통해 128개 동시 접속을 허용할경우 1000개를 동시 접속 진행할경우
nginx에서 1차적으로 초당 200개 요청, 대기 설정 1000명중 800명 대기 진행해서 평균 응답속도가 5-7초
가 나오는것을 확인했습니다.

예시 **1000개 동시 접속**

Worker : 8개
thread : 1개 woker x 16개



- 초당 200개 요청 제한
- 대기 800명(1000명 설정)

128개 동시 접속 허용(16x8)

동시 접속 테스트시 평균 속도(5~7초 소요)

Type	Name	# Requests	# Fails	Median (ms)	95%ile (ms)	99%ile (ms)	Average (ms)	Min (ms)	Max (ms)	Average size (bytes)	Current RPS	Current Failures/s
GET	/api/test	19621	0	2100	15000	117000	5037.63	6	146624	47	467.2	0
	Aggregated	19621	0	2100	15000	117000	5037.63	6	146624	47	467.2	0