주요 용도	대규모 모델 훈련·추론, 대용량 벡터 연산, HPC 워크로드	경량 추론(특히 FP16 mixed precision), 엣지 AI, 소형 배포 환경
장점	- 대규모 파라미터 모델(수백억~수조) 훈련 가능 - 높은 메모리 대역폭으로 대규모 배치 처리 - MIG(Matrix Instance GPU)로 자원 분할 기능 지원	- 전력 효율이 뛰어나며, 소규모 모델 추론 비용 대비 성능 우수 - 저전력 데이터센터나 클라우드 환경에 적합
단점	- 초기 투자 비용이 매우 높음 - 유지보수 및 전력 소비가 큼	- 메모리 제한(16GB)으로 대형 모델 구동 불가 - 연산 성능이 대형 모델 훈련/추론에 부족

Q. DO 솔루션에서 Local LLM 사용할 때 GPU 서버 구축시 A100이 최소사양인 이유는?

A. A100은 메모리 용량과 대역폭이 훨씬 크고 연산 성능이 뛰어나므로 대규모 LLM 추론과 다중 처리에 유리합니다..

대량의 벡터 연산과 복잡한 RAG 연산을 원활히 수행하려면 대용량 메모리와 높은 연산량이 필수적이므로, DO 솔루션 최소 사양으로 A100 이상의 GPU가 요구됩니다.

— 2025.09.13 ~ 업데이트 — 팔란티어 관련

9. 팔란티어 관련

9.1 팔란티어 솔루션 별 설명

Q. 팔란티어 솔루션으로는 무엇이 있고 각각 어떤 분야에서 활용되며 특징은 무엇인가요?

A. Gotham : 정부 기관용 위협 분석 및 정보 통합 도구

정부·국방 분야에서 사용되는 데이터 분석 플랫폼.

테러리스트 추적, 군사 작전 지원, 사이버 보안 등에 활용.

미국 정보기관 (예: CIA, FBI), 국방부 (DoD) 등에서 사용

Foundry : 기업의 데이터를 통합·시각화해 의사결정을 지원하는 플랫폼