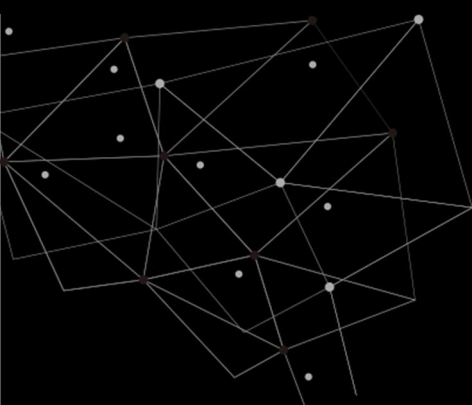
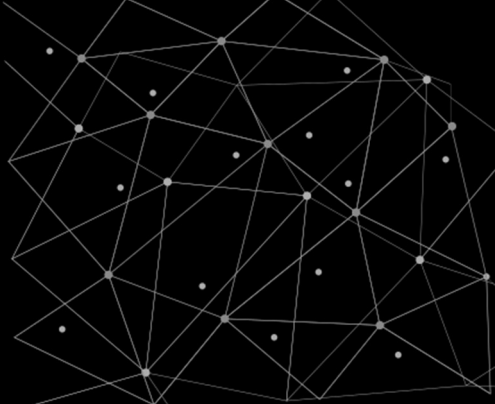


2026 오픈소스 AI 활용 인재 양성 특화형 과정

Gen AI on EKS/K8s

인퍼런스 최적화

Project Guide



1

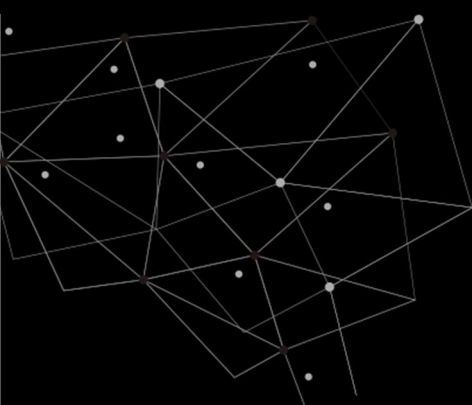
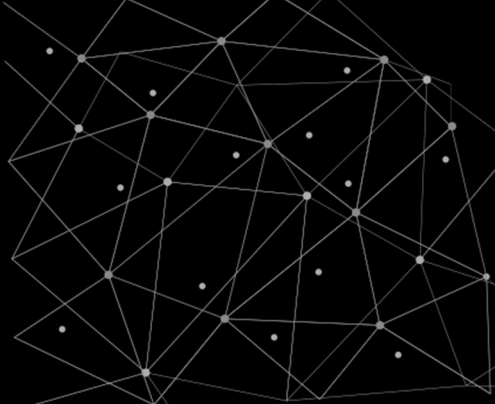
과정 개요

활용 기술 / 난이도 / 활용 언어 등



① 과정 개요

- ✓ **과정명** : GenAI on EKS/K8s 인퍼런스 인프라 최적화
- ✓ **난이도** : ★★★★★☆
- ✓ **활용 언어** : Terraform(HCL), YAML, Python, Shell Script
- ✓ **활용 기술** : Kubernetes (Amazon EKS), vLLM, SGLang, LiteLLM, EFA, Mountpoint for S3, Karpenter



2

과정 소개

과정 상세 소개



2

과정 소개



프로젝트 세부 내용

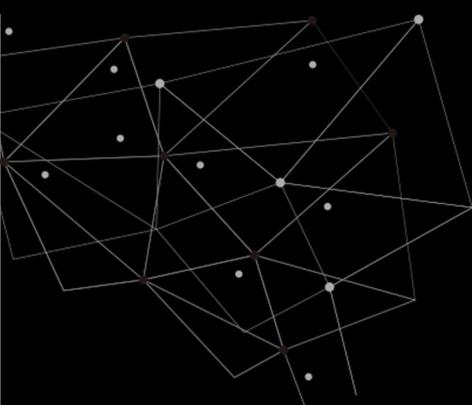
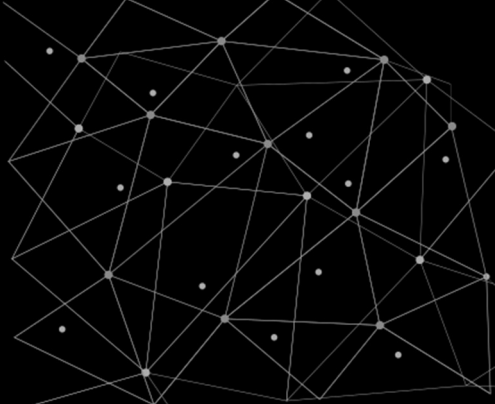
[프로젝트 목표]

생성형 AI 워크로드(vLLM, SGLang 등)가 Kubernetes 위에서 운영될 때, 오픈소스 인퍼런스 스택과 실제 AWS 인프라(EFA, Mountpoint for S3 등)위에서 오픈소스로 구동할 때, 네트워크·스토리지 계층을 포함한 다양한 부분에서 성능을 최적화하는 방안을 직접 체험 및 그 결과를 aws-samples/sample-genai-on-eks-starter-kit 오픈소스에 기여 (MIT-0 라이선스)

[운영 계획]

- 1주차: EKS 클러스터 배포, GenAI Starter Kit 설치, Terraform 인프라 분석
- 2주차: vLLM/SGLang 배포, LiteLLM AI Gateway 구성, 모델 서빙 파이프라인 구축
- 3주차: EFA 설정, 멀티노드 분산 추론 네트워크 튜닝, RDMA 고속 통신 구현
- 4주차: Mountpoint for S3 구성, 모델 가중치 로딩 최적화, FSx for Lustre 캐싱
- 5주차: Langfuse/Phoenix 추론 모니터링, 성능 메트릭 수집 및 대시보드 구성
- 6주차: 최적화 전/후 Latency·Throughput 비교, 병목 분석
- 7~8주차: 개선 사항 문서화, PR 작성·제출, 코드 리뷰 반영 후 Merge

※ 세부 일정 및 내용은 상황에 따라 변경될 수 있음



3

멘토 소개

담당 멘토 소개



3

멘토 소개



정영준 멘토
(리드멘토)

✓ 주요 경력

현재

- Amazon Web Services (AWS), Sr. Specialist Solutions Architect, Containers & ML Infra (APJC)

이전

- Red Hat, Solutions Architect
- Zespro, Solution Architect / Senior Software Engineer

✓ 주요 활동

- aws-samples/sample-genai-on-eks-starter-kit 리포지터리 메인테이너
- The Engineering Playbook 메인테이너([링크](#))
- AWS Korea Tech Blog 기고: [링크1](#), [링크2](#) 등
- AWS re:Invent 2020 및 다수 주요 행사 발표: [링크1](#), [링크2](#) 등

3

멘토 소개



최영락 멘토

✓ 주요 경력

현재

- Amazon Web Services (AWS), Containers GTM Specialist

이전

- Microsoft (2018. 11. - 2023. 8.)
- Fusetools Korea (2017. 4. - 2018. 4.)

✓ 주요 활동

- Kubernetes SIG-Docs, Localization Subproject Approver & Reviewer
- OpenStack I18n SIG Lead
- 국내 OpenStack Korea Group 대표 (2017-2019)