**Air Cloud Documentation (Updated Korean Version) /** 25.08.20 Updated  
  
[기능 추가 요청건]  
- 메인 우측 (ENG/KOR) 언어 선택 지원  
-[Home | AirCloud](https://integrated-dev.aieev.cloud:3002/ko) 해당 페이지 우측 Air Cloud 띄고 로고 반영해주세요 ^^; 혹은 다른 로고 있어요 전달드릴게요

**1. Overview – Air Cloud란**

**Air Cloud는 대규모 AI 추론을 안정적으로 처리할 수 있는 GPU 클라우드 플랫폼입니다.** 실시간 AI 서비스부터 비용 최적화가 필요한 워크로드까지, 사용자 목적에 맞는 두 가지 옵션을 제공합니다.

* **Air Cloud**: 분산형 GPU 인프라를 기반으로, 검증된 네트워크를 통해 **효율적이고 합리적인 비용**을 제공합니다.
* **Air Cloud+**: AIEEV가 직접 운영하는 전용 GPU 노드로, **높은 가용성·성능·안정성**을 보장합니다.

전통적인 전용 GPU 클라우드가 **실서비스(프로덕션) 환경에서 겪는 한계**를 해결하기 위해 설계된 Air Cloud를 통해 복잡한 설정이나 비용 부담없이 시작해보세요.

지금 바로 [계정을 만들고 시작해보세요](https://integrated-dev.aieev.cloud:3002/ko)  
-> 링크로 연결되게 부탁드립니다.( [Home | AirCloud](https://integrated-dev.aieev.cloud:3002/ko))

**A. 배포 방식 선택**

* **Air Container** : AWS, GitHub, Docker Hub, Google 등 퍼블릭 · 프라이빗 레포지토리에서 컨테이너 이미지를 손쉽게 배포할 수 있습니다. 이미지 주소와 보안 토큰만 입력하면 즉시 실행 가능합니다.
* **Air Model** (‘25년 9월 출시 예정): 서버리스 기반 초 단위 과금 모델로, 빠른 AI 추론을 위한 환경을 제공합니다. 모델을 자동으로 불러오고 REST 엔드포인트를 구성하며, 프로덕션 환경을 위한 오토스케일링까지 지원합니다.
* **Air API** (‘25년 9월 출시 예정(?) : Whisper, LLaMA 3.3 8B, Solar Mini, D.N.A와 같은 오픈소스 모델을 활용하여 OpenAI 호환 REST API 엔드포인트를 제공합니다.

**B. 인프라 설정**

* **Air Cloud** : 한국 내 1만여 대 이상의 GPU가 연결된 분산 네트워크로, 검증된 GPU 파트너 네트워크를 통해 운영됩니다. 이 중 1,000대 이상의 GPU를 사전 확보해 두었으며, 작업에 따라 동적으로 할당되어 업계 최고 수준의 비용 경쟁력을 제공합니다.
* **Air Cloud+** : AI추론 전용으로 설계된 AIEEV가 직접 운영하는GPU 서버 높은 가용성과 서비스 안정성, 엔터프라이즈급 보안을 제공합니다.

**C. 미션**

AIEEV는 성능, 안정성, 사용성을 결코 타협하지 않으면서 누구나 GPU AI 클라우드 플랫폼을 쉽고 합리적인 비용으로 활용할 수 있도록 설계했습니다. 개인 사용자부터 기업까지, 모두가 AI와 확장형 인프라의 잠재력을 최대한 활용할 수 있도록 지원합니다.

**도움이 필요하신가요?**   
사용안내서에서 확인할 수 없는 내용은,   
Discord([Discord](https://discord.com/channels/912829806415085598/1187492973148115076/threads/1349672733168107541)), 이메일(contact@aieev.com)을 통해 문의해주세요  
-> 누르면 링크 바로 연결될 수 있도록 부탁드립니다.

**D. 진행 가이드**

* Air Cloud 튜토리얼 실행하기
* 계정에 사용 금액 충전하기

**2. 시작하기**

**Air Cloud에 오신 것을 환영합니다.**   
이 가이드를 순차적으로 따라 하시면 5분 이내에 첫 번째 배포를 완료할 수 있습니다.

리소스를 실행하려면 결제 계정과 최소 1개 이상의 프로젝트가 필요합니다.   
Air Cloud의 콘솔(관리 화면)에서 단계별 설정 안내 절차를 통해 이 과정을 순차적으로 진행할 수 있습니다.

**A. 계정 생성하기**

**텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

그림 1. 가입 및 로그인

먼저 Air Cloud 계정을 생성해 배포를 관리하고 사용할 수 있도록 하세요.  
1)**GitHub**, 2)**Google**, 3)**이메일 주소**를 사용하여 가입할 수 있습니다.   
회원 가입이 완료되면 로그인하여 플랫폼을 바로 이용할 수 있습니다.

**B. 크레딧 충전하기**

신규 사용자는 가입 시 **$20의 무료 크레딧**이 제공됩니다. 리소스를 배포하기 위해서는 계정에 크레딧이 충전된 상태여야 합니다. 계좌 이체 및 신용카드 결제를 통해 크레딧 충전이 가능합니다.

대량 크레딧 충전을 원하시면 결제 이전 [sales@aieev.com](mailto:sales@aieev.com)으로 이메일을 보내주세요.   
담당 팀이 영업일 1일 이내 연락을 드립니다.   
대량 크레딧 구매 및 전용 리소스 예약 시 추가 할인을 안내드립니다.

**C. 프로젝트 생성하기**

텍스트, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 2. 프로젝트 생성

계정에 크레딧을 충전한 후, 워크로드를 배포할 프로젝트를 생성합니다. 하나의 프로젝트 내에서 여러 개의 컨테이너, 모델, API를 배포해 사용자의 AI 서비스 아키텍처에 맞게 구성할 수 있습니다.

**D. 조직(Organization) 관리하기**

개인 계정을 생성하거나, 팀 멤버가 초대한 계정으로 참여할 수 있습니다.

멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 3. Organization Members 페이지

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

그림 4. Member 초대 창

* **조직 생성**: GitHub, Google, 이메일 주소를 통해 계정을 생성하고 조직을 만듭니다.
* **팀원 초대**: 관리자(Owner)는 [Members] 메뉴에서 이메일로 팀원을 초대할 수 있습니다. 단, 초대받은 팀원은 **초대받은 이메일 주소과 동일한 주소로 가입을 완료 후 초**대를 수락해야 조직 및 프로젝트에 참여할 수 있습니다.
* **권한 유형**: 관리자(Owner)는 팀 설정, 멤버 초대 및 삭제, 프로젝트 제거 등을 추가로 관리할 수 있으며, 일반 사용자(Member)는 배포 및 모니터링 도구에만 접근할 수 있습니다.

**참고**: 초대를 받은 사용자는 초대 수락 전 모든 계정에서 로그아웃해야 오류 *(e.g., 404 page)* 없이 정상적으로 초대가 처리됩니다.

**E. 결제 정보**

* **Air Container** 및 **Air Model** 은 **시간 단위**로 과금됩니다. 자세한 정보는 [**Billing FAQ]**를 참고해주세요.
* **결제 수단** : 국내외 **계좌 이체 또는 신용카드 결제로 충전 및 결제를 진행할 수 있습니다.**

**3. Air Container**

Air Container는 컨테이너 기반의 AI 서비스를 빠르고 효율적으로 배포할 수 있도록 지원합니다.

**[기본 설정]**

멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 5. General 정보 입력

스크린샷, 텍스트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 6. Image 정보 입력

컨테이너 생성 시 다음 항목들을 입력해야 합니다.

* **이름(Name)**: 컨테이너를 식별하기 위한 사용자 정의 이름
* **카테고리(Category)**: 컨테이너가 제공하는 서비스 유형 (향후 Playground 연동에 사용됩니다)
* **컨테이너 이미지(Container Image)**: 컨테이너의 Docker 이미지 URL
* **레지스트리 제공자(Registry Provider)**: 이미지가 호스팅된 레지스트리 (예: GitHub Container Registry, Docker Hub 등)
* **레지스트리 사용자 이름 및 비밀번호(선택 사항)**: 프라이빗 레지스트리에서 이미지를 불러올 경우 입력 필요

**[고급 설정]**

**스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

그림 7. 고급 설정

Advanced 섹션에서는 다음과 같은 추가 옵션을 설정할 수 있습니다:

* **시작 명령어(Start Command)**: 컨테이너 이미지의 기본 시작 명령어를 재정의
* **포트(Port)**: 컨테이너에서 기본 노출할 포트 재정의
* **헬스 체크 URL(Health Check URL)**: 컨테이너의 상태를 확인할 헬스 체크 URL 지정. (예: /api/health).
* **환경 변수(Environment Variables)**: 애플리케이션 실행에 필요한 key-value 형태 설정 (예: DB 인증 정보, API 키 등)

**[배포 완료]**

텍스트, 소프트웨어, 멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 8. 배포된 Container 정보

컨테이너가 배포 즉시 실행되며, **API 엔드포인트** 또는 **서비스 URL**이 자동으로 제공됩니다.

* 생성된 컨테이너는 좌측 목록에 표시됩니다.
* 목록에서 항목을 선택하면 우측 패널에 상세 정보가 표시됩니다.
* 우측 패널에서 컨테이너 설정을 수정할 수 있습니다.
* “대시보드(Dashboard)" 버튼을 클릭하면, 컨테이너 상태 및 활동 로그를 확인할 수 있는 상세 관리 페이지로 이동합니다.

**[API 요청]**

컨테이너 상태가 실행 중(RUNNING)일 때, 노출된 엔드포인트를 통해 AI 추론 API에 접근할 수 있습니다.

요청 시 호스트 주소 부분을 해당 컨테이너의 엔드포인트 URL로 교체해야 합니다..

**API 요청 예시:**

curl --request POST \

--url ${ENDPOINT\_URL}/api/v1/chat/completions \

--header "Accept: application/json" \

--header "Authorization: Bearer ${YOUR\_API\_KEY}" \

--header "Content-Type: application/json" \

--data '

{

"messages": [

{

"role": "system",

"content": "You are a helpful assistant."

},

{

"role": "user",

"content": "Write a haiku about recursion in programming."

}

]

}'

Air Container는 컨테이너화된 AI 서비스를 빠르고 효율적으로 배포할 수 있도록 합니다. 이용 방법은 다음과 같습니다:

스크린샷, 멀티미디어 소프트웨어, 그래픽 소프트웨어, 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **프로젝트 선택**: 프로젝트 대시보드(Project Dashboard)로 이동한 뒤, 초기설정 과정 중에 생성한 조직과 프로젝트를 선택합니다.

스크린샷, 멀티미디어 소프트웨어, 소프트웨어, 그래픽 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **이용 서비스 선택**: "컨테이너(Container) 서비스를 선택합니다.

멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **기본 설정**: 컨테이너 이름을 입력하고, 클라우드 유형(Air Cloud / Air Cloud+)을 선택한 뒤, 복제(Replica) 수를 지정합니다.

스크린샷, 텍스트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **컨테이너 이미지 설정**: 이미지 경로(Docker Hub, AWS ECR 등), 인증 정보, 레지스트리 설정을 입력합니다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. **리소스 설정**:
   * **일반 모드(General Mode)**: 인스턴스 유형을 선택합니다 (예: RTX 4070, 4090).
   * **오토스케일 모드(Autoscaling Mode)**: 최소 및 최대 복제 수를 설정합니다 (1~30개). 그 이상의 높은 수치를 설정할 경우 관리자 승인이 추가적으로 필요합니다. 필요시 [contact@aieev.com](mailto:contact@aieev.com)로 문의해주세요.

스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. **고급 설정:** 모델 시작 명령어, 포트, 헬스 체크 URL, 환경 변수를 설정합니다.

스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **검토 및 배포**: 사용자 설정 내용을 최종 확인한 후 컨테이너를 배포합니다. 이미지는 최초 다운로드 이후 캐시되어, 첫 실행 이후부터는 컨테이너 시작 시간을 단축할 수 있습니다.

**4. 모니터링 및 API 테스트**

* **단일 요청 테스트**: Postman 등 API 테스트 도구를 사용하여 vLLM 컨테이너의 엔드포인트를 확인할 수 있습니다.
* **오토스케일링 테스트**: Locust를 사용해 부하를 시뮬레이션하고 레플리카 확장 동작을 검증할 수 있습니다.
* **사용량 지표**: 레플리카 수, 요청 처리율, 시스템 응답 시간 등을 모니터링할 수 있습니다.
* **실행 로그(Runtime Logs)**:
  + 각 컨테이너 인스턴스 및 레플리카별로 로그가 표시됩니다.
  + Air Cloud 대시보드에서 실시간 및 과거 로그를 확인할 수 있습니다.
  + 로그는 최신 순(내림차순)으로 정렬됩니다.
  + 시간 범위, 컨테이너, 인스턴스 ID 기준으로 필터링할 수 있습니다.
  + 로그에는 시작 명령어 출력, 헬스 체크 상태, 오류 메시지, 모델 서버의 stdout/stderr 정보가 포함됩니다.
  + 컨테이너가 실패한 경우에도, 로그는 일정 기간 동안 보존되어 디버깅에 활용할 수 있습니다.
* **설정 변경**: 엔드포인트 설정은 컨테이너가 정지된 상태에서만 수정할 수 있습니다.

**5. 문제 해결**

실제 사용자 사용 과정 중 **자주 발생하는 문제 사례와 해결 방법**입니다.  
아래를 확인하여 순서대로 처리할 수 있습니다.

**[자주 발생하는 HTTP 오류]**

* **401 Unauthorized**: 인증 실패의 경우, API 키 또는 접근 권한 정보를 다시 확인해 주세요.
* **404 Not Found**: 요청한 컨테이너 ID 또는 엔드포인트 경로가 잘못된 경우입니다. 컨테이너 ID 또는 라우트를 다시 확인해 주세요.
* **500 Internal Server Error**: 내부 서버 오류가 발생할 경우, 잠시 후 다시 시도하거나 Air Cloud 지원팀([contact@aieev.com](mailto:contact@aieev.com))으로 문의해 주세요.

**[A]. API 키 인증 실패시**

* **증상**: API 요청 인증에 실패함
* **원인**: 만료되었거나 잘못된 API 키 사용
* **해결 방법**:
  + API 키가 유효한지 확인합니다 (만료 여부 확인)
  + Air Cloud 대시보드에서 새로운 API 키를 발급받아 사용

**[B]. GPU 할당 오류 발생시**

* **증상**: 컨테이너 시작 시 GPU가 감지되지 않음
* **원인**: 노드 불안정성 또는 드라이버 불일치
* **해결 방법**:
  + 새로운 인스턴스로 재배포
  + 사용 중인 컨테이너 런타임과 드라이버의 호환성 확인

**[C]. 네트워크 연결 실패**

* **증상**: 컨테이너가 외부 엔드포인트에 연결하지 못함
* **원인**: 포트 설정 오류, DNS 문제, 방화벽 차단
* **해결 방법**:
  + 컨테이너가 올바른 포트를 노출하고 서비스가 수신 중인지 확인
  + 컨테이너 내부에서 curl 명령어로 외부 연결 테스트
  + 보안 그룹 설정에서 아웃바운드 트래픽이 차단되지 않았는지 확인

**[D]. 헬스 체크 실패 / 초기화 단계에서의 멈춤**

* **증상**: 배포가 헬스 체크 단계에서 멈춤
* **원인**: /health 경로에서 HTTP 200 OK 응답을 반환하지 않음
* **해결 방법**:
  + 올바른 헬스 체크 url 및 포트가 노출되어 있는지 확인
  + 로그를 확인하여 에러 메시지나 모델 로딩 지연 여부 파악

**[E]. 오토스케일링 미작동**

* **증상**: 높은 부하에도 불구하고 확장이 발생하지 않음
* **원인**: 스케일링 임계값이 너무 높거나, 레플리카 상태가 비정상
* **해결 방법**:
  + CPU 임계값, 지연 시간 등 오토스케일링 파라미터 조정
  + 모든 레플리카가 헬스 체크를 정상 통과하는지 확인

**[F]. 502/503 게이트웨이 오류**

* **증상**: API 응답이 서버 오류(502/503)를 반환
* **원인**: 애플리케이션이 예상 포트에 바인딩되지 않았거나 시작 중 충돌 발생
* **해결 방법**:
  + 서비스가 0.0.0.0:<PORT>에 정상적으로 실행 중인지 확인
  + 모델 로딩 시간이 길 경우 헬스 체크 타임아웃을 늘려서 설정

**6. 문의하기**

추가 담당 지원이 필요하신 경우, 아래 담당자에게 연락해 주세요:

* **김현경 매니저**: [hk.kim@aieev.com](mailto:hk.kim@aieev.com) / +82-10-9888-4696
* **박병수 매니저**: [bspark@aieev.com](mailto:bspark@aieev.com) / +82-10-4102-4050