



인공지능 중심 산업융합 집적단지 조성사업

# HPC-AI 공용인프라 구축사업

## 포탈 사용자 교육

2022. 11. 3



# INDEX

1. 회원가입/대시보드
2. 볼륨
3. 데이터셋
4. 컨테이너
5. 잡스케줄링
6. 모델/모델배포
7. 노트북
8. 강의
9. 모니터링
10. 기술문의

1.

# 회원가입 /대시보드

# 1.1. 그룹관리자의 그룹사용자 초대

그룹  
Home / Management / 그룹

테스트그룹

+ 하위그룹 생성 수정 삭제

그룹 그룹의 기본정보와 할당 정책을 확인하고 관리 할 수 있습니다.

그룹 기본 정보

그룹 아이디	testgroup	그룹명	테스트그룹
그룹구분	회사	초대키	iqoL09iKcE5QLGghNnog 그룹 키 복사
단체(회사/학교) 명		팀(부서/학과) 명	

그룹 가입링크 메일전송

초대할 사용자의 이메일을 입력하세요.

master@aifrica.co.kr

입력 후 심표 또는 엔터를 클릭하세요. 여러명일 경우 엔터로 구분해서 입력하세요.

전송

테스트그룹 그룹으로 초대합니다.

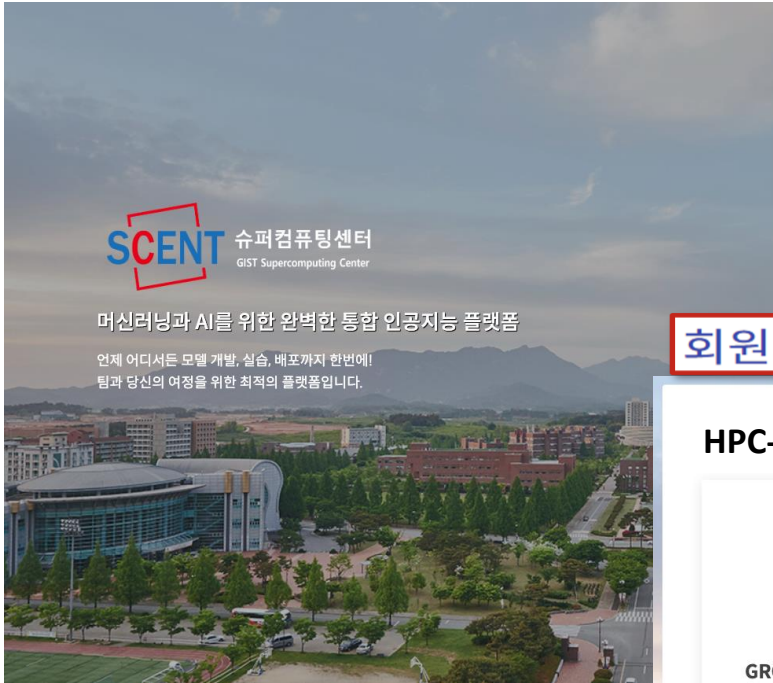
안녕하세요.  
그룹관리자님이 귀하를 이 그룹으로 초대했습니다.

테스트그룹 가입하기

※ 본 이메일은 발신 전용입니다. 궁금하신 사항은 로컬 시스템 운영자나 아래 연락처로 문의해주시기 바랍니다.

## 1.2. 회원가입

### 1) 회원가입유형 선택



로그인

username

Password

로그인

gist 로그인

회원가입

### HPC-AI 서비스 신청 회원가입



#### GROUP 관리자

새로운 그룹(기관, 기업, 학교 등)을 생성하고 그룹을 관리하는 사용자

- 새로운 그룹을 생성
- 그룹사용자 초대와 관리
- 그룹이 사용 가능한 자원(GPU, CPU) 관리

선택하기

#### GROUP 사용자

이미 생성되어 있는 그룹에 소속된 사용자로 서비스를 이용하는 경우

- 그룹으로부터 초대 받아 그룹 가입키를 가지고 있는 경우
- 그룹 사용기간 다양한 리소스 공유 가능
- 그룹 관리자가 자원 할당 후 사용

선택하기



#### 개인 사용자

그룹에 포함되지 않는 일반 사용자로 서비스를 이용하는 경우

- 자원(GPU, CPU) 배정 후 사용
- 지타를 체험하고 싶은 사용자

선택하기

이미 가입되어 있다면 로그인 하세요.

## 1.2. 회원가입

### 2) 그룹가입키로 그룹사용자 가입

### HPC-AI 서비스 포탈 회원가입

그룹가입키를 입력하세요.

**관리자에게 가입키를 받아 가입을 진행합니다.**

그룹가입키를 입력하세요.

이미 가입되어 있다면  하세요.



### HPC-AI 서비스 포탈 회원가입

회원정보를 입력하세요.

**이용 약관**

\* 이용 약관에 동의해주세요. [보기](#)

**개인정보처리방침**

\* 개인정보처리방침에 동의해주세요. [보기](#)

**회원 정보**

그룹  에이프리카

\* 사용자 ID  사용자 아이디는 영문자, 숫자, '-' 로만 가능합니다.

\* 이름

\* 비밀번호

\* 비밀번호 확인

\* 전화번호

\* 이메일

\* 이메일 알림여부

**부가 정보**

단체(회사/학교)명  단체명을 입력하세요.

팀(부서/학과)명  팀명을 입력하세요.

이미 가입되어 있다면  하세요.

## 1) 그룹사용자/일반사용자 대시보드

Welcome HPC-AI 서비스 포탈

**작업환경 생성**

- 데이터셋 추가
- 블록 추가
- 노트북 추가
- 잡 생성

**작업환경 생성**

다양한 라이브러리와 도구를 탑재한 빌트인 이미지들을 통해 최적의 컨테이너 기반 ML 작업 및 실험 환경을 구성할 수 있습니다. 지금 바로 컨테이너를 생성하여 작업 환경을 준비하고 실험을 시작할 수 있습니다.

[컨테이너 생성](#) → 바로가기

[컨테이너에 대해서 더 알아보기](#)

**Document**

[처음 사용자](#)  
처음 이용하는 사용자들 위한 도움말

[리소스 활용하기](#)  
사용이 익숙하고 다양한 기능을 활용하고 싶은 사용자들 위한 도움말

HPC-AI 서비스 포탈 사용 안내

등록된 공지사항이 없습니다.

**강의** 가장 인기 있는 강의 → 더보기

- test course
- PPT 강의
- 파이썬과 머신러닝(초급)-test

**데이터셋** 가장 인기 있는 데이터셋 → 더보기

- DeepLearningExamples from github

**노트북** 가장 인기 있는 노트북 → 더보기

- 데이터 증강
- PDF 강의 일니다. -test
- recent

**GPU 일간 사용량** 분 단위 사용량 보정

**현재 리소스 사용 현황** 리소스 사용 현황

Resource	Type	User	Group
Intel(R) Xeon(R) Silver 4208 CPU @ 2.10GHz	CPU	0 of 12.8 Cores	0 of 14.4 Cores
		0 of 5.5 Gi	0 of 6.4 Gi
NVIDIA-P102-100	GPU	3 of 8	37.5%
		3 of 8	37.5%

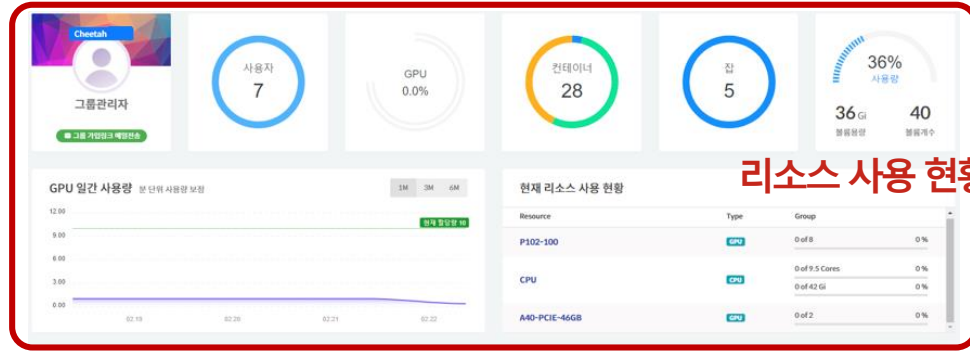
**내 최근 활동**

- 구동중인 컨테이너 3
- 잡 현황(계획/전체) 0/2
- 블록 현황 320 Gi
- 블록개수 29
- ubuntu20.04-cuda11.3-cudnn8 15시간 전
- ubuntu20.04-cuda11.1-cudnn8 하루 전
- 통합 테스트 하루 전
- 데이터 증강 37분 전
- 질문을 등록한다. 한 시간 전
- scheduler 변경 테스트 15시간 전
- caost local model 17시간 전
- recent 하루 전
- caost local model deployment 하루 전
- gpu-workspaceubuntu18.04-cuda10.0-cudnn7 하루 전
- flower photo 하루 전
- qwer flower 5일 전
- flower\_photos 5일 전

최근 활동

# 1.3. 대시보드

## 2) 그룹관리자 대시보드



**사용자 승인 대기 목록** (0)

사용자 승인 대기 목록이 없습니다.

**사용중인 컨테이너 목록** (0)

컨테이너 목록이 없습니다.

**최근 작업 목록** (5)

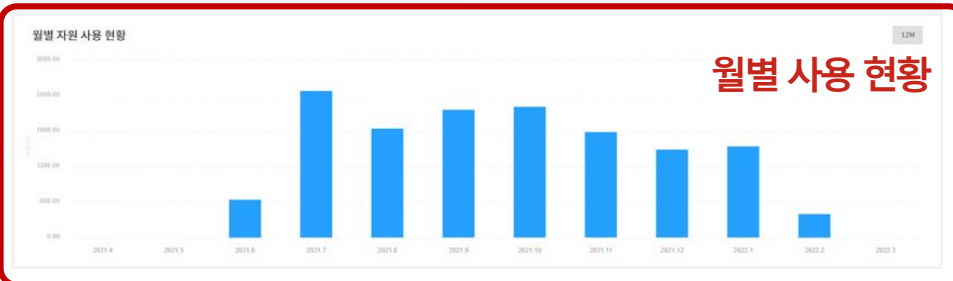
이름	상태	리소스	작업명
kepcoco demo job	종료됨	1GB	kepcoco demo
Test	종료됨	1GB	kepcoco demo
Test	종료됨	1GB	kepcoco demo
Test	종료됨	1GB	kepcoco demo
Test	종료됨	1GB	kepcoco demo

**블로그 목록** (0)

이름	리소스	블로그명
kepcoco demo	1GB	kepcoco demo
kepcoco demo	1GB	kepcoco demo
kepcoco demo	1GB	kepcoco demo
kepcoco demo	1GB	kepcoco demo

**사용자 이미지 목록** (3)

이름	리소스	이미지명
NFS	25.3 GB	NFS
hifan sample	25.5 GB	hifan sample
hifan sample_customer	25.5 GB	hifan sample





2.

# 보름

## 2.1. 볼륨 유형

**볼륨**  
Home / Management / 볼륨




컨테이너에서 자동으로 생성되는 컨테이너 별 Home 볼륨과 사용자 정의 볼륨을 사용할 수 있습니다.  
생성된 볼륨은 같은 그룹원끼리 공유할 수 있습니다. 또한 SFTP등을 사용하지 않고 파일브라우저를 활용하여 웹UI를 통해 손쉽게 파일을 업로드 할 수 있습니다.

도움말 보기

볼륨 이름  조회

내 볼륨    나에게 공유된 볼륨

총 3건 (1/1)

볼륨 이름	상태	볼륨 공유 구분	스토리지 타입	파일브라우저/SFTP	등록자	등록일
 customvolume 5 GB \$USER_HOME/customvolume	사용 가능	사용자 공유	WEKA		그룹유저 groupuser	몇 초 전
 demo 컨테이너의 홈 볼륨 10 GB \$USER_HOME	사용 가능	사용자 공유	CEPH		수수 SOOSOO	5달 전
 demo 5 GB \$USER_HOME/demo	사용 가능	사용자 공유	CEPH		수수 SOOSOO	5달 전



**Home 볼륨**

컨테이너 생성시 기본으로 제공되는 볼륨으로  
컨테이너 생성과 함께 자동 생성 됩니다.



**사용자 볼륨**

필요에 따라 만들어서 다른 컨테이너나 Job에서  
공유할 수 있는 볼륨입니다.

## 2.2. 사용자 볼륨 생성

### 볼륨

- \* 볼륨 이름
- \* 볼륨 경로
- \* 설명
- \* 신청 용량
- \* 볼륨 공유 구분

공유 대상 사용자

목록

Volume

볼륨 경로는 생성 후 변경 불가능하니 신중히 결정해 주세요.  
볼륨 경로는 영문 대소문자, 숫자, -, \_로만 최소 4자리에서 최대 24자리까지 사용 가능합니다.

볼륨의 간략한 설명 및 용도

0 GB 5 GB 10 GB

그룹 유저의 볼륨 MAX쿼터는 그룹관리자가 지정한 용량으로 설정됩니다.

사용자 공유

- 사용자 공유 - 일반 컨테이너에서 사용 가능하며, 생성한 볼륨을 사용자에게 공유할 수 있다.
- 그룹 공유 - 그룹컨테이너 생성 시에만 사용할 수 있으며, 그룹컨테이너 생성 시 사용자들에게 읽기 전용으로 보인다.

전체 선택

(group-user)

저장


jupyter

Files Running Clusters

Select items to perform actions on them.

0  /

Volume



## 2.3. 파일브라우저

**1** 파일탐색기/SFTP 시작

**2** 파일탐색기/SFTP 중지

**3** 파일 업로드

File Browser

manager

.....

로그인

demo

주소: 주의 요함 | 183.111.96.75:32143/files/

내 파일

새로운 폴더

새로운 파일

설정

로그아웃

폴더가 비어 있습니다...

File Browser 2.0.16  
도움말

대용량 파일 또는 100개 이상의 다량의 파일을 한번에 업로드 하는 경우, 안정적인 파일 전송을 위해 파일브라우저보다는 SFTP 사용을 권장

3.

# 데이터셋

# 3.1. 데이터셋 기본유형 업로드

## 데이터셋

Home / Management / 데이터셋

데이터를 탐색, 분석 및 공유할 수 있습니다. 공유된 데이터셋은 차타 컨테이너에서 사용할 수 있습니다.

[도움말 보기](#) [+ 데이터셋 업로드](#)

Search datasets

검색결과 2 개 (0.005초)

[공개 데이터셋](#) [그룹 데이터셋](#) [내 데이터셋](#)

1

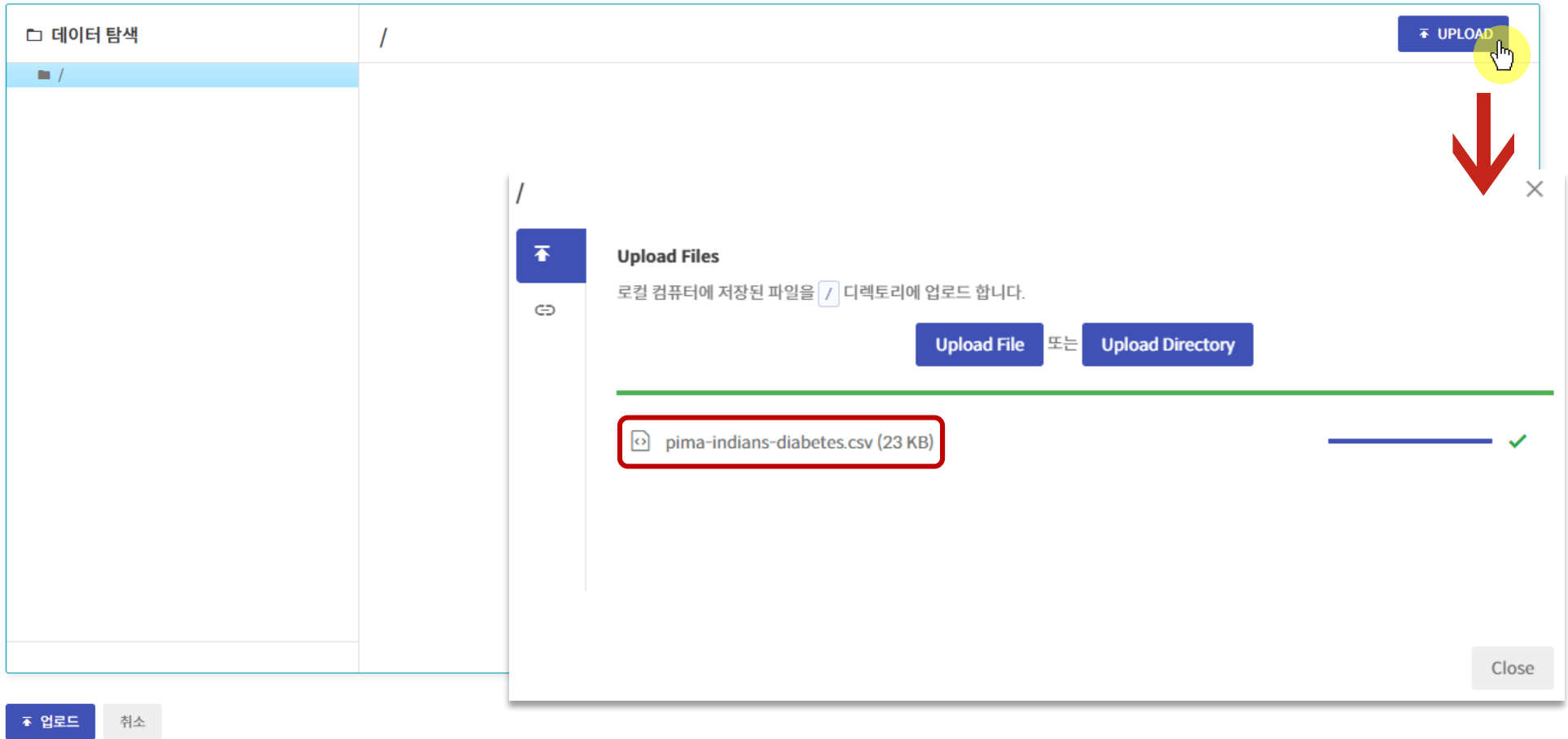
### 데이터셋 업로드

데이터를 탐색, 분석 및 공유할 수 있습니다. 공유된 데이터셋은 차타 컨테이너에서 사용할 수 있습니다.

#### 데이터셋 정보

- \* 공유 유형: 전체 공유
- \* 데이터셋 제목: 피마인디언
- \* 라이선스: CC0: Public Domain
- \* 데이터셋 설명: 당뇨병 예측
- \* 태그: 태그 입력 후 엔터  
태그는 최대 10개까지 입력 가능합니다. 공백을 허용하지 않습니다( \_로 치환되어 저장)
- 댓글 활성화:  댓글 활성화
- 데이터셋 유형: 기본

### 3.1. 데이터셋 기본유형 업로드



## 3.2. 데이터셋 GITHUB 유형 업로드

데이터셋 유형

GitHub Repository

**i** GitHub Repository를 사용하여 데이터셋을 생성합니다.

Preview

- ▶ .github
- ▶ .gitignore
- ▶ .gitmodules
- ▶ CUDA-**데이터셋**
- ▶ DGLPy
- ▶ Faster
- ▶ Kaldi
- ▶ MxNet

**기본 정보**

DeepLearningExamples from github

**PUBLIC** **CC0: Public Domain** groupuser · 몇 초 전 · Views 0 · Downloads 0 · deeplearning examples

**GitHub** <https://github.com/NVIDIA/DeepLearningExamples/tree/63cbbdc2588f7925a4e5cc30bc6e0cb51d9c166a>

DeepLearningExamples from github

Uploading

요청한 데이터셋 파일을 업로딩 중입니다.  
잠시만 기다려주세요.



4.

# 컨테이너

## 4.2. 컨테이너 생성

### ☑ 컨테이너 정보

- 컨테이너 이름
- 설명

100자 이하의 컨테이너 이름

컨테이너의 설명 및 용도

### 1 기본 정보 입력

### ☒ 자원

- 자원

✓ GPU

• CPU

### 2 GPU/CPU 자원 선택

ON\_PREMISE\_GPU  
○ P102-100

NVIDIA, 3200 core, @ 5.059Mi  
VCPU 3.2 core, Memory 63.0 Gi

#### ② 내 GPU 사용 현황



#### ② 노드의 할당 가능한 GPU 현황



● 사용중 ● 사용 중인 GPU에 할당 ● 할당됨 ● 대기중

0 개

1 개

### ☞ 이미지

- 이미지 선택

✓ 빌트인 이미지

• 사용자 이미지

### 3 빌트인 이미지와 사용자 이미지 중 선택

cheetah/gpu-workspace:ubuntu18.04-cuda10.1-cudnn7

cheetah/gpu-workspace:ubuntu18.04-cuda10.1-cudnn7  
cheetah/gpu-workspace:ubuntu18.04-cuda10.0-cudnn7  
cheetah/gpu-workspace:ubuntu16.04-cuda9.0-cudnn7  
cheetah/gpu-workspace:ubuntu18.04-cuda11.0-cudnn8  
cheetah/cpu-workspace:r-verse3.6.2  
cheetah/cpu-workspace:r-geospatial3.6.2  
cheetah/matlab:r2020a-lnu-addon

### ☒ 가상 환경(커널) 옵션

- 가상 환경

# 4.1. 컨테이너 생성\_사용자 이미지

이미지

• 이미지 선택

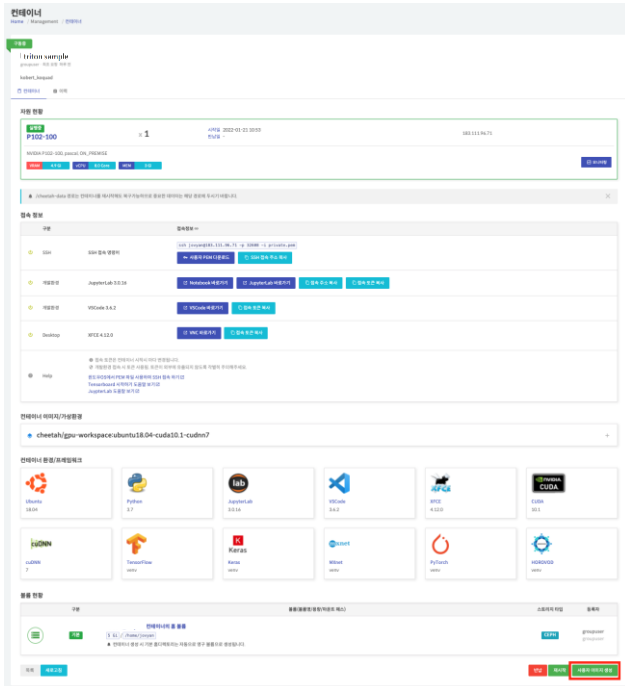
• 빌트인 이미지

✓ 사용자 이미지

➕ 사용자 이미지 선택

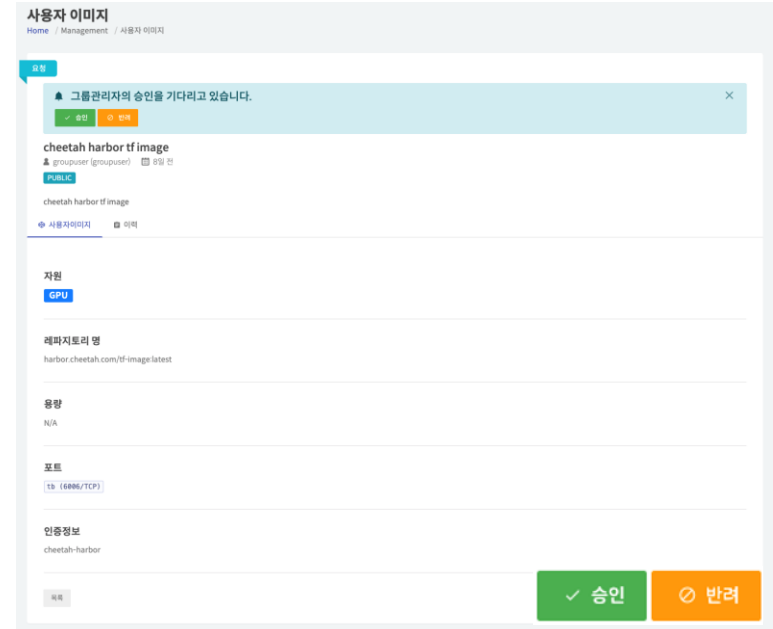
## 커스텀 이미지

구동중인 컨테이너의 특정 시점에 컨테이너를 이미지로 저장하고 볼륨 정보를 스냅샷으로 저장할 수 있습니다.



## 외부 이미지

개인 또는 공개 레파지토리의 이미지 관리자에게 요청 후 승인이 되면 외부이미지사용이 가능합니다.



# 4.1. 컨테이너 생성

## 컨테이너 환경

구성 환경

A grid of 12 icons representing different container environments and tools. Each icon includes a logo and text: Ubuntu 18.04, Python 3.7, JupyterLab 2.2.8, VSCode 3.6.2, XFCE 4.12.0, NVIDIA CUDA 10.1, cuDNN 7, TensorFlow venv, Keras venv, MXnet venv, PyTorch venv, and HOROVOD venv.

## 볼륨

볼륨

### 4 기본 홈볼륨 외 사용자 볼륨/데이터셋 추가 가능

A screenshot of a volume management interface. At the top left, there is a green button labeled '볼륨 연결 추가' (Add Volume Connection) with a hand icon. Below it, a table shows volume details:

볼륨(볼륨명/용량/마운트 패스)	스토리지 타입	등록자	Action
기본 Home 1 Gi / \$USER_HOME	CEPH	데모사용자 user01	삭제

저장 취소



A screenshot of the 'Volumes' panel. It shows two volume entries:

- CEPH test**: 5 Gi, test. Includes a '추가' (Add) button.
- NFS demo**: 1 Gi, demo. Includes a '추가' (Add) button.

A screenshot of the 'Datasets' panel. It shows two dataset entries:

- TIME -Image Dataset-Classification**: PUBLIC. Includes tags: classification, computervision, education, +2 More. Includes a '추가' (Add) button.
- Breast Cancer Wisconsin(Diagnostic)**: PUBLIC. Includes tags: data, disease. Includes a '추가' (Add) button.

## 4.2. 컨테이너 실행

Cheetah/cpu-workspace-tensorflow

컨테이너 환경/프레임워크

- Ubuntu 18.04
- Python 3.6
- JupyterLab 2.1.0
- TensorFlow venv
- Keras venv

볼륨 현황

구분	볼륨 이름	마운트 경로	마운트 타입	볼륨 용량
기본	home	/home/jovyan ⚠ 컨테이너 생성 시 기본 홈디렉토리는 자동으로 영구 볼륨으로 생성됩니다.	ON_PREMISE	1 Gi
공유	test	/home/jovyan/test	ON_PREMISE	1 Gi

목록 새로고침

컨테이너

Home / Management / 컨테이너

생성 수정 삭제

구동중

자원 현황

실용성	사작일	반납일	비고
NVIDIA GeForce RTX-2080-Ti x 1	2022-03-14 12:06	-	183.111.68.76

NVIDIA NVIDIA GeForce RTX-2080-Ti turing\_ON\_PREMISE

VRAM 10.8 Gi vCPU 8.0 Core MEM 62 Gi

이 상태

⚠ /cheetah-data 경로는 컨테이너를 재시작해도 복구가능하므로 중요한 데이터는 해당 경로에 두시기 바랍니다.

## 4.2. 컨테이너 실행

자원 현황

**실행중** **NVIDIA-GeForce-RTX-2080-Ti** x 1      시작일 2022-03-14 12:06      183.111.68.76  
 받남일 -

NVIDIA NVIDIA-GeForce-RTX-2080-Ti, turing, ON\_PREMISE

**VRAM** 10.8 Gi    **vCPU** 8.0 Core    **MEM** 62 Gi    [모니터링](#)

🔔 /cheetah-data 경로는 컨테이너를 재시작해도 복구가능하므로 중요한 데이터는 해당 경로에 두시기 바랍니다.

접속 정보

구분	접속정보
SSH	SSH 접속 명령어 <div style="border: 2px solid red; padding: 5px;"> <span>ssh jovyan@183.111.68.76 -p 30421 -i private.pem</span>  <span>사용자 PEM 다운로드</span>    <span>SSH 접속 주소 복사</span> </div>
개발환경	JupyterLab 3.0.16 <span>Notebook 바로가기</span> <span>JupyterLab 바로가기</span> <span>접속 주소 복사</span> <span style="border: 2px solid red; padding: 2px;">접속 토큰 복사</span>
개발환경	VSCode 3.11.1 <span>VSCode 바로가기</span> <span>접속 토큰 복사</span>
Desktop	XFCE 4.12.0 <span>VNC 바로가기</span> <span>접속 토큰 복사</span>
Help	<ul style="list-style-type: none"> <li>접속 토큰은 컨테이너 시작시 마다 변경됩니다.</li> <li>개발환경 접속 시 토큰 사용됨. 토큰이 외부에 유출되지 않도록 각별히 주의해주세요.</li> </ul> <p>원도우OS에서 PEM 파일 사용하여 SSH 접속하기 <a href="#">☞</a>                      Tensorboard 시작하기 도움말 보기 <a href="#">☞</a>                      JupyterLab 도움말 보기 <a href="#">☞</a></p>

**1** 접속 토큰 복사

**3** 복사한 토큰 입력

jupyter

**3** 복사한 토큰 입력

Password or token:

**Token authentication is enabled**

If no password has been configured, you need to open the notebook server with its login token in the URL, or paste it above. This requirement will be lifted if you [enable a password](#).

The command:

```
jupyter notebook list
```

# 4.3. 컨테이너 반납/삭제

**컨테이너 반납시에는 영구 볼륨(홈, 사용자 정의) 이외 정보는 재사용이 불가능합니다.**

**반납**

구분	볼륨(볼륨명/용량/마운트 패스)	스토리지 타입	등록자
기본	test 컨테이너의 홈 볼륨 10 Gi / /home/jovyan 컨테이너 생성 시 기본 홈디렉토리는 자동으로 영구 볼륨으로 생성됩니다.	CEPH	배정아 manager

**홈볼륨도 같이 삭제되어 복구가 불가능하므로 중요한 데이터는 사용자 볼륨을 추가하여 사용하시길 권장 드립니다.**

**삭제**

컨테이너가 성공적으로 반납되었습니다.

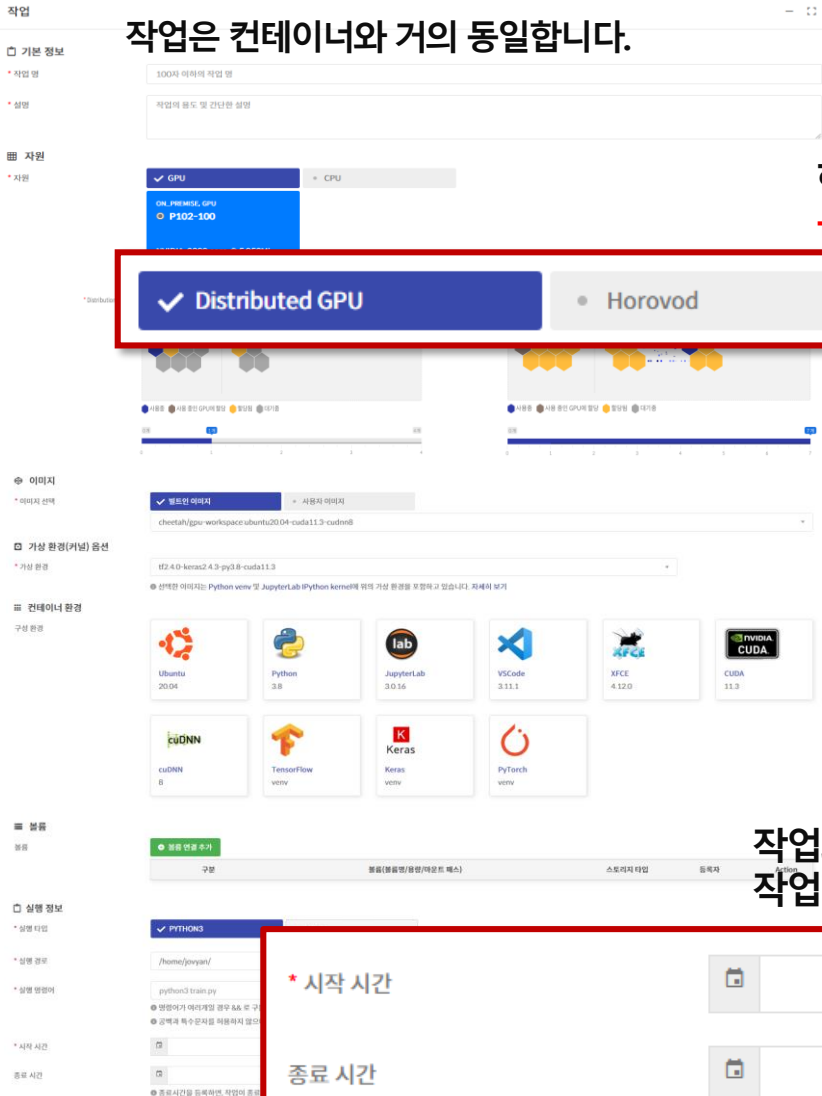
구분	볼륨 이름	마운트 경로	마운트 타입	볼륨 용량
기본	home	/home/jovyan 컨테이너 생성 시 기본 홈디렉토리는 자동으로 영구 볼륨으로 생성됩니다.	ON_PREMISE	1 Gi
공유	test	/home/jovyan/test	ON_PREMISE	1 Gi

5.

# 잡스케줄링



# 5.1. 잡스케줄링 생성



작업은 컨테이너와 거의 동일합니다.

하지만 GPU 자원을 선택할 때 분산된 GPU와 멀티 클러스터링 할 수 있는 Horovod 중 선택할 수 있습니다.

\* 물리적으로 다른 서버에 있는 GPU를 묶어서 한 서버에서 사용하는 것처럼 실행 가능

작업의 시작 시간과 종료 시간을 설정할 수 있지만 종료시간은 필수가 아닙니다. 작업이 종료되지 않아도 해당 시간에 강제종료되기 때문입니다.

\* 시작 시간

종료 시간

🕒 종료시간을 등록하면, 작업이 종료되지 않아도 해당 시간에 강제종료됩니다.

# 5.1. 잡스케줄링 생성

**작업**  
Home / Management / 작업

**Fasion\_MNIST\_estimator\_job**  
수수 (soosoo), 2021-11-24 16:21

**테스트**  
python3 Fasion\_MNIST\_estimator\_job.py

**컨테이너 정보**

주 자원	NVIDIA-GeForce-RTX-2080-Ti x 1
△ Distribution Type	Distributed GPU
○ Image	cheetah/gpu-workspace:ubuntu20.04-cuda11.3-cudnn8
⊙ Kernel	tf2.4.0-keras2.4.3-py3.8-cuda11.3
≡ 볼륨	/home/jovyan/ 10GB /home/jovyan/soosoo 5GB

**작업 실행 정보**

실행 타입	PYTHON3
실행 경로	/home/jovyan/soosoo
실행 명령어	python3 Fasion_MNIST_estimator_job.py
예약된 런타임	2021-11-24 16:21:57 (예약된 런타임이 없습니다)
런타임	확정

**작업이력**  
새로고침

테스트

**작업이력**

#	파드명	단계	상태	시작일	종료일	로그	상세	모니터링
1	job-736f6f736f6f-kadiy5-44h6p	테스트	실행중	2021-11-24 16:21:57		로그	상세	보기

새로고침

**작업이력**

#	파드명	단계	상태	시작일	종료일	로그	상세	모니터링
1	job-736f6f736f6f-kadiy5-44h6p	테스트	종료(성공)	2021-11-24 16:21:57	2021-11-24 16:23:02	로그	상세	보기

새로고침

잡스케줄링 요청 테스트 수정 삭제

테스트가 정상적으로 성공하면 예약된 시간에 작업이 실행됩니다.

## 5.2. 잡스케줄링 확인

### 스케줄

[Home](#) / [Management](#) / 스케줄

계획된 작업들의 일정을 한 눈에 확인할 수 있습니다.

< October 2021 >

Month Week Day List

SUN	MON	TUE	WED	THU	FRI	SAT
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3 12a Demo Job - groupu	4	5	6

6.

# 모델/모델배포

# 6.1 모델 등록

모델 등록 내 모델을 직접 업로드 할 수 있습니다. 컨테이너에서 모델 생성 방법은 도움말을 참조해 주세요.

**모델 기본정보**

- 모델 공유 유형: **비공유**, 그룹 공유, 전체 공유
- 모델 이름:
- 모델 레파지토리 이름:   
● 모델 레파지토리 이름은 배포시 URL 구분값으로 사용됩니다. 생성 후 변경 불가능하니 신중히 결정해 주세요.  
모델 레파지토리 이름은 공백 없이 숫자, 영문 소문자, 영문 대문자, '\_', '-'만 가능하며 최소 4자리에서 최대 64자리까지 사용 가능합니다.
- 모델 설명:
- 모델 유형:
- 모델 파일 업로드: **사용 가능한 모델에 대한 자세한 도움말은 [여기서 확인 하세요.](#)**  
  
Drag and drop a model ZIP file here or click

## 사용 가능한 Model 레이아웃

모델 서빙은 다양한 머신러닝 프레임워크에 대응하기 위하여 [Nvidia Triton Inference Server](#)를 사용하여 구동됩니다.


[Triton Inference Server](#) (이하 Triton)는 CPU와 GPU 모두에 최적화된 클라우드 추론 서비스를 제공합니다. Triton은 HTTP/REST 또는 gRPC 엔드포인트 제공을 통해 추론 서비스를 제공하며, 원격 클라이언트가 서버에서 관리되는 모델에 대한 추론을 요청할 수 있도록 합니다. Edge 배포를 위해 전체 기능을 확장할 수 있는 C API를 제공하며 다양한 클라이언 제작을 위해 Python/C++ 클라이언트 라이브러리를 제공합니다.

```
<model-zip-file>/
[config.pbtxt]
[output-labels-file] ...]
<version>/
<model-definition-file>
...
saved_model.zip
├── 1
│   └── saved_model.pb
│       └── variables
│           ├── variables.data-00000-of-00002
│           ├── variables.data-00001-of-00002
│           └── variables.index
└── config.pbtxt
```

이때 모델 파일은 **Triton Inference Server**에서 제시하는 레이아웃에 맞춰서 준비하셔야 합니다.

## 6.1 모델 등록

노트북 파일 업로드

  
Drag and drop a notebook file here or click

모델 요약(JSON)

```
1 summary here json type
  {
    "accuracy": 0,
    "loss": 0
  }
```

모델 파라미터

```
1 {
  "param": ""
}
```

플러그인 파일

↕ ❌ +

↕ 업로드 취소

## 6.2 모델 배포

### 모델

Home / Management / 모델

모델 기본정보

**demo**  
수수 (soosoo) 5달 전  
PUBLIC TensorFlow

목록

**모델 배포** 수정 삭제

### 버전 정보

모델 버전 별 조회

배포 상태 없음 + 버전 추가

버전	상세	생성일	Actions
1	4 Model files	5달 전	+ 보기


#### Navigation

- 모델 기본정보
- 모델 버전 정보
- 모델 파라미터
- 노트북
- 댓글

## 6.2 모델 배포

### 모델 배포 관리

배포된 모델은 HTTP, GRPC등의 엔드포인트를 활용하여 서빙됩니다.

- 기본 정보**
  - 배포 이름: 100글자 이하의 배포 이름
- 모델 선택**
  - 모델 선택: [모델 선택](#)  demo
- 모델 배포 공유**
  - 모델 공유 유형: 비공유
- 모델 배포 파라미터**
  - Instance 유형: RTX 2080 Ti (NVIDIA-GeForce-RTX-2080-Ti x 1, 1000 Milli core, 16384 Mi)  
인스턴스는 사용자의 할당량 안에서 선택할 수 있으나 생성시에 시스템의 잔여량에 따라 실행되지 않을 수 있습니다.
  - Image: nvcr.io/nvidia/tritonserver:20.08-py3
  - Environment variables:

```
{
  "KEY1":"VALUE1",
  "KEY2":"VALUE2"
}
```

Example:

```
{ "LD_PRELOAD": "/plugins/liblayerplugin.so" }
```

[저장](#) [취소](#)



```
~ curl -v \  
  -H "Content-Type: application/json" \  
  -H "model-deployment-sn: 888" \  
  -X GET http://193.123.96.86:8000/v2/health/live  
Note: Unnecessary use of -X or --request, GET is already inferred.  
* Trying 193.123.96.86...  
* TCP_NODELAY set  
* Connected to 193.123.96.86 (193.123.96.86) port 8000 (#0)  
> GET /v2/health/live HTTP/1.1  
> Host: 193.123.96.86:8000  
> User-Agent: curl/7.64.1  
> Accept: */*  
> Content-Type: application/json  
> model-deployment-sn: 888  
>  
< HTTP/1.1 200 OK  
< content-length: 0  
< content-type: text/plain  
< x-envoy-upstream-service-time: 4  
< date: Thu, 17 Dec 2020 05:32:31 GMT  
< server: istio-envoy  
<  
* Connection #0 to host 193.123.96.86 left intact  
* Closing connection 0
```

7.

노트북

# 7.1 노트북 등록

## 노트북

Home / Management / 노트북

실험 중인 컨테이너에서 업로드하거나 직접 업로드한 Jupyter Notebook을 다른 사용자에게 자유롭게 공유하고 의견을 나눌 수 있습니다. 공유된 노트북은 치타 컨테이너에서 자유롭게 추가하여 사용할 수 있습니다.

도움말 보기

+ 노트북 업로드

### 노트북 업로드 정보

공유 유형: 전체 공유

노트북 제목: 노트북 제목을 간략하게 입력하세요.

노트북 설명: 노트북 설명을 간략하게 입력하세요.


노트북 파일 업로드: Drag and drop a Jupyter Notebook file here or click


첫번째 노트북 파일은 자동으로 Version 1으로 저장됩니다.

태그: 태그 입력 후 엔터  
태그는 최대 10개까지 입력 가능합니다. 공백을 허용하지 않습니다(로 치환되어 저장)

댓글 활성화 여부:  댓글 활성화

업로드 취소





**Keras 전처리 레이어 사용하기**

참고: 이 섹션에서 소개하는 Keras 전처리 레이어는 현재 실험적 단계입니다.

**크기 및 배율 조정하기**

전처리 레이어를 사용하여 이미지를 알려진 모양으로 크기 조정하고 픽셀 값의 배율을 조정할 수 있습니다.

```
[7]: IMG_SIZE = 180

resize_and_rescale = tf.keras.Sequential([
    layers.experimental.preprocessing.Resizing(IMG_SIZE, IMG_SIZE),
    layers.experimental.preprocessing.Rescaling(1./255)
])
```

**Navigation**

- 노트북 기본정보
- 노트북
- 데이터 증강
- 개요
- 설정
- 데이터세트 다운로드
- Keras 전처리 레이어 사용하기
- tf.image 사용하기
- 다음 단계
- 댓글 1

## 7.2. 컨테이너로 복사

### 노트북 기본정보

#### [LightGBM Classifier] 고객 만족 데이터셋을 이용한 고객 만족 예측 분석

**PUBLIC** 시스템관리자 · 7달 전 · Views 44 · Downloads 7 · `grid_searchcv_parameter_tuning`  
`lightgbm_santander_custsat` `plot_importance_for_feature_importance` `sklearn_eda`  
`xgboost_classifier_customer_satisfaction`

Version 1 ▾

캐글의 산탄데르 고객 만족 데이터셋의 고객 만족 여부를 XGBoost와 LightGBM을 활용해 예측. 산탄데르 고객 만족 예측 분석은 370개 피처로 클래스 레이블 명은 Target이며 1이면 불만 고객, 0은 만족. 데이터 전처리와 auc 성능평가를 기준으로 GridSearchCV로 최적 파라미터를 탐색하여 XGBClassifier로 이전 0.8445에서 0.8463으로 개선. LightGBM Classifier로 GridSearchCV로 최적파라미터를 통해 ROC AUC 0.8417.

목록

... 더보기 ▾

+ 노트북 버전 추가

📄 컨테이너로 복사

↓ 다운로드

Jupyter

Files Running Clusters

Select items to perform actions on them.

Upload New

0 / Notebooks

Name Last Modified File size

LightGBM-Classifier]-고객-만족-데이터셋을-이용한-고객-만족-예측-분석.ipynb Running 몇 초 전 541 KB

Jupyter [LightGBM-Classifier]-고객-만족-데이터셋을-이용한-고객-만족-예측-분석 Last Checkpoint: 한 시간 전 (autosaved) Python 3 (pykernel)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel) Memory: 465.7 MB / 63.5 GB

### XGboost & LightGBM Classifier: Santander Customer Satisfaction

캐글의 산탄데르 고객 만족 데이터셋의 고객 만족 여부를 XGBoost와 LightGBM을 활용해 예측. 산탄데르 고객 만족 예측 분석은 370개의 피처로 클래스 레이블 명은 Target이며, 1이면 불만 고객, 0은 만족

데이터는 <https://www.kaggle.com/santander-customer-satisfaction/data> 에서 다운로드 받을 수 있음. train.csv를 train\_santander.csv로 변경

데이터 전처리 XGBoost는 사이킷런 레퍼를 이용

```
In [12]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib

cust_df = pd.read_csv("./train_santander.csv", encoding='latin-1')
print('dataset shape:', cust_df.shape)
cust_df.head(3)
```

```
Out[12]:
```

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_comer_ult3	imp_op_var
0	1	2	23	0.0	0.0	0.0	0.0	0.0	0.0
1	3	2	34	0.0	0.0	0.0	0.0	0.0	0.0
2	4	2	23	0.0	0.0	0.0	0.0	0.0	0.0

8.

강의



# 8.1. 강의

## 강의

Home / Management / 강의

차타에서 제공하는 머신러닝과 프로그래밍 언어등의 기본 강의를 활용할 수 있습니다.

도움말 보기

+ 강의 등록

### 강의 기본정보

#### Python과 머신러닝

**PUBLIC** 시스템관리자 · 몇 초 전 · Views 161 · deeplearning keras machinelearning python tensorflow

#### [강의개요]





본 수업은 Python을 이용하여 머신러닝의 기초 개념을 알고 실습수업을 통해 이해 및 적용하는 것이 목표입니다. 중간과서 이전에는 기초 Python 강의와 머신러닝과 딥러닝에 대한 개념을 확립하는 이론 위주의 수업으로 진행될 예정입니다. 이후에는 Tensorflow와 Keras를 사용하여 직접 코드를 작성해 앞서 공부한 이론을 적용할 수 있도록 진행할 것입니다. 기본적으로 머신러닝과 딥러닝은 수학에 기본을 두고 있기 때문에 기초 행렬 및 함수에 대한 이해가 있는 학생들이 비교적 쉽게 접근할 수 있습니다.

[교재] PPT, Jupyter Notebook

[과제] 일일 퀴즈(Python)강의에 해당(1~3주), 개인과제

목록

### 강의 목차

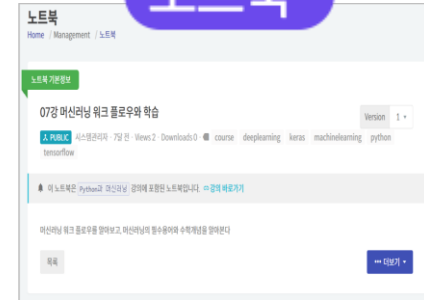
- 1 피이션과 머신러닝 강의교안 - 주차별 강의내용 PDF  ... 더보기
- 2 오리엔테이션 PDF  ... 더보기
- 3 01강 Python 기본 맛보기 Jupyter notebook 인기 많은 언어이 Python에 대하여 소개하고, 주석사용법, Print, Input, variable, type, string, boolean, list, tuple, dictionary 등을 설명합니다.  ... 더보기
- 4 02강 Python의 문자열 리스트, 튜플, 딕셔너리 Jupyter notebook Python의 문자열 리스트와 튜플, 딕셔너리에 대해서 자세히 알아본다.  ... 더보기



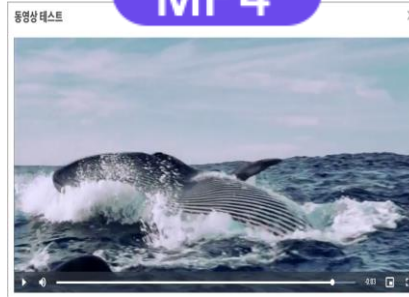
PDF



노트북



MP4



Markdown

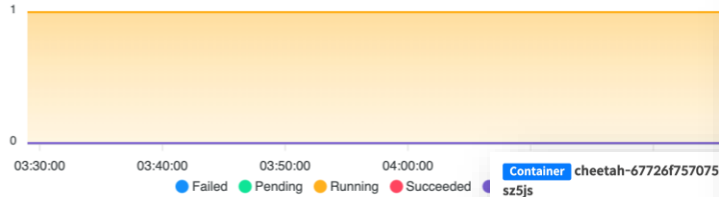
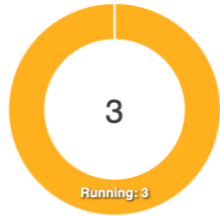


9.

# 모니터링

# 9.1. 모니터링

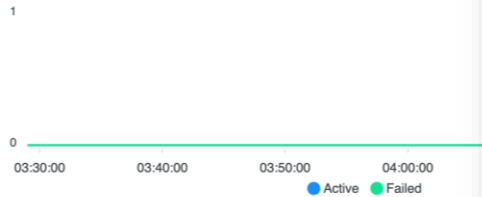
## Container



#	Container	Pod	Node	사용자
1	Demo container	cheetah-67726f757075736572-mxeifz-8455dbf7d5-sz5js	dev-gpu01	group1
2	custom image test	cheetah-67726f757075736572-om4q7i-5c7b86bf64-4k4	dev-gpu-01	group1
3	dataset test	cheetah-67726f757075736572-zekrmn-7f5dc8f866-tj4w	dev-gpu-02	group1

## Job

0



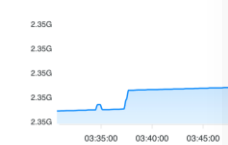
#	Job	Pod	사용자	상태
실행 중인 Job이 없습니다.				

Container cheetah-67726f757075736572-sz5js

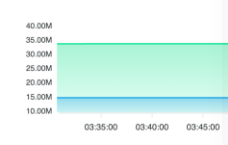
CPU Requests



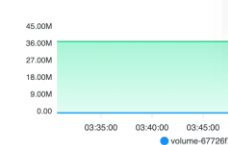
Memory Requests



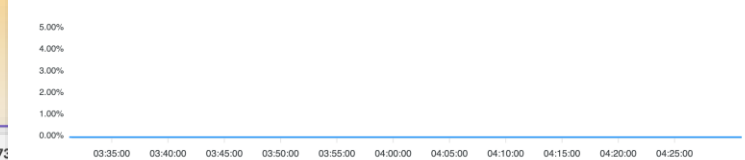
Network Usage



Disk Usage



GPU Usage



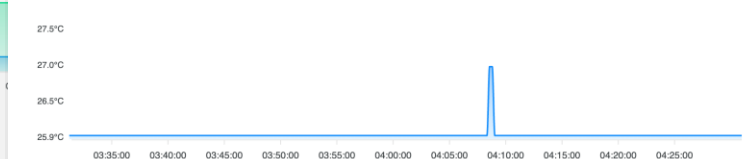
GPU Memory Usage



GPU Power Usage



GPU Temperature





10.

# 기술문의

## 10.1. 기술문의

HPC-AI 서비스 포탈 웹 하단에 고정 되어있는 고객지원을 클릭하시면 Help Desk로 이동합니다.

### 문의 등록

이메일 주소 \*

고객사명 \*

계정명 \*

연락처 \*

• 개인정보 수집에 동의한것으로 간주하며 문의 답변 외의 용도로 사용되지 않습니다.

문의 유형 \*

지타 사용 시 문제점이나 장애 문의는 "기술 지원" 유형을 선택해주시고, 기타 다른 문의는 제공해 드리는 적절한 옵션을 선택해 주세요.

제품 영역 \*

지타 서비스 중 장애/문의 내용과 관련 있는 유형을 선택해 주세요.

제목 \*

설명 \*

Africa의 직접 고객을 대상으로 문의를 지원하고 있습니다. 직접 고객이 아닌 유상 기술 지원을 요청하실 경우 "영업 문의"로 등록하여 주십시오.