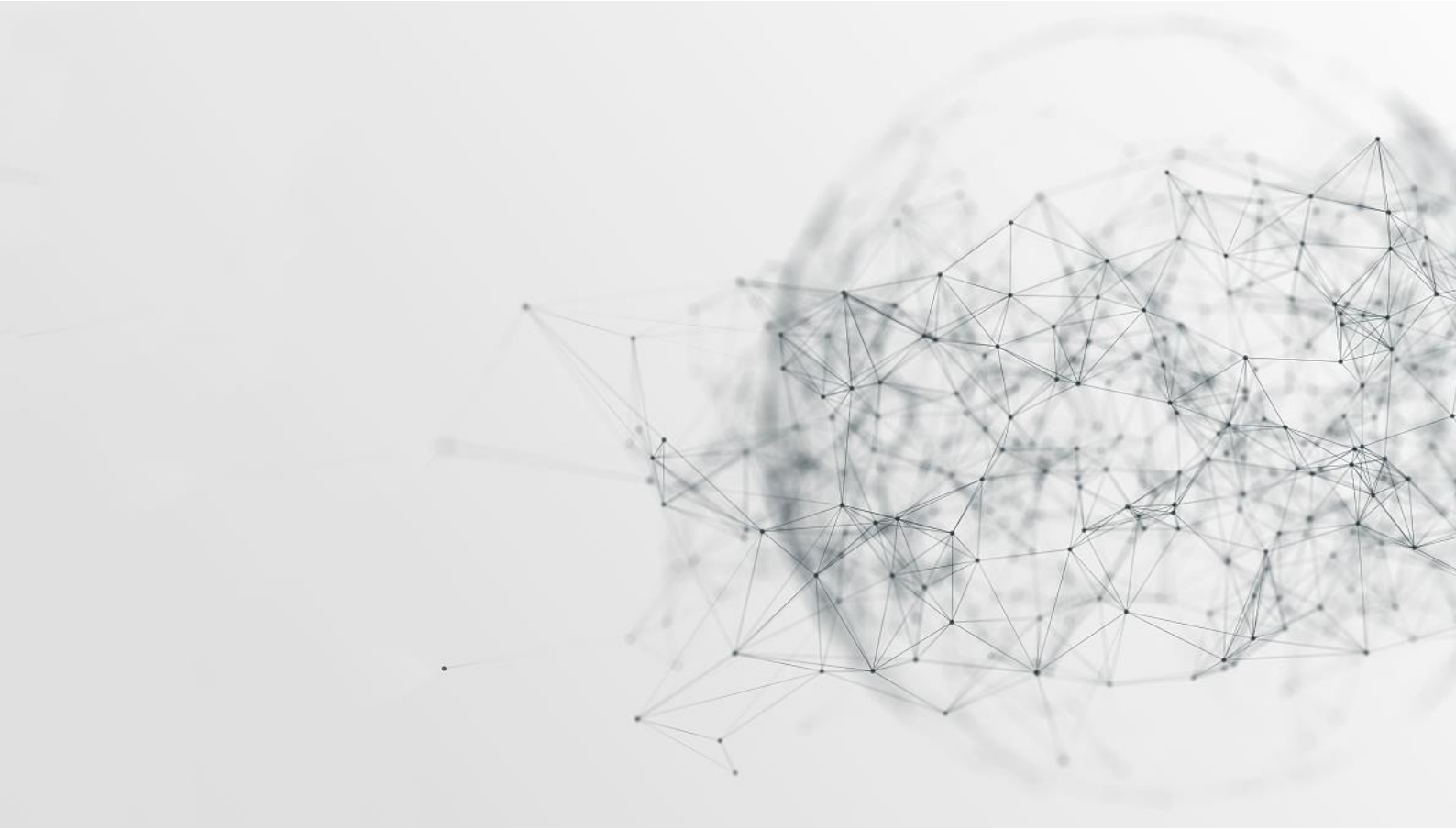
APULIS 依瞳科技

Democratize Enterprise AI



Apulis AI Platform

User Manual

Version: **1.5.0**

Release Date: **04 02 2021**

Apulis Technology (Shenzhen) Co., LTD

www.apulis.com

目录

1 User Guide

Through this document, you will learn how to train deep neural networks on Apulis AI platform, along with ways to manage your training jobs and monitor the executing status of your jobs or the running status of cluster nodes.

The target audience for this document includes deep learning algorithm engineer, AI application developers, researchers, students and all other AI platform intended users.

2 Product Introduction

2.1 Product description

Apulis AI platform aims to provide an end-to-end deep learning development platform for users in various industries, which empowers users to carry out deep learning research in the fastest and most time-efficient way that greatly reduces development cost and improves productivity. For small and medium size enterprises, the platform can also help to lower the technical threshold and reduce the costs of building private cloud infrastructure.

The platform provides development environments such as model training, code development, model management, data management, online reasoning services, resource monitoring, virtual clusters, etc., so that AI developers can quickly build artificial intelligence development environments and develop AI applications. A pre-warning module is built based on the monitoring module to automatically notify the administrator of platform exceptions to improve the pre-warning efficiency and safety of the platform.

2.2 Glossary

Terms, abbreviations	Explanation
Tensorflow	TensorFlow is a large-scale deep learning framework

	on distributed systems. It has good portability, can run on mobile devices, supports distributed multi-machine multi-card training, and supports various deep learning models.
PyTorch	PyTorch is developed by the Facebook AI team. Unlike TensorFlow, PyTorch uses eager execution mode and is python native, makes it easy to learn and integrate with other Python packages.
MindSpore	MindSpore is a new open-source deep learning training/inference framework that could be used for mobile, edge and cloud scenarios. MindSpore is designed to provide development experience with friendly design and efficient execution for the data scientists and algorithmic engineers, native support for Ascend AI processor, and software hardware co-optimization. At the meantime MindSpore as a global AI open-source community, aims to further advance the development and enrichment of the AI software/hardware application ecosystem.
Kubernetes	K8S for short is an open-source container-orchestration system for automating application deployment, scaling and management. It aims to provide a platform for automating deployment, scaling, and operations of allocation containers across clusters of hosts.
engine	A common technical framework, kernel or environment in model development and training.
Code development	Model training jobs in Expert System.
VC	Virtual Cluster, the physical Cluster of all AI Processor group management, each group is a Virtual Cluster.
Device type	Computing resources that can be used for model training, such as Nvidia GPU, Atlas NPU.
Device quantity	The number of resources that can be allocated to training jobs.
Image	A collection of files required to perform model training.
Jupyter	Jupyter Lab is an interactive web computing environment that supports multiple programming languages. It is convenient to write algorithm code, submit model training jobs, and debug.
SSH	SSH or Secure Shell is a cryptographic network protocol for operating network services securely over an unsecured network.
TensorBoard	TensorBoard is a visualization tool, which can be used

	to show the network diagram, the index changes of tensors, the distribution of tensors etc. When training the network, user can set different parameters (such as weight W, bias B, number of convolutional layers, number of full connection layers, etc.), and TensorBoard can be used to intuitively select parameters.

2.3 Function introduction

This section mainly introduces the functions and usage of the platform, including resource allocation, user and authentication management, code development, virtual cluster management, model training, and data management. Users submit model training jobs through the web interface. You can view the status of the running job on the overview page, and on the resource monitoring page, you can view the real-time resource usage and the log output of the training job, etc.; through cluster status monitoring, you can view the resource usage of the entire cluster and monitor the status of the physical nodes.

3 Operation Instruction

3.1 Access Platform

Users can access Apulis AI platform through browsers. Please note that Chrome 87.0.4280.88 (official version) is recommended. The default entrance page is [Login page] as shown in Figure 1.

Figure 1: Login page

3.1.1 Register/Login

Click [Register Account] on the login page to jump to the registration page. You need to enter username, nickname and password to finish registration. After registration, you will be redirected to the [Login Page]. After login to the platform, new users will see the following prompt, "Sorry, you do not have access to the current page". Please contact your administrator to grant you relevant authorities and connect your account with virtual clusters (VC). Only after this can you access the full functionality of the platform.

If you already have an account, you can just enter the account and password on the [Login Page] and click the Login button to login.

Figure 2: Register Page

3.2 User Resource Manage

3.2.1 Operations Administrator Login to User manage system

Enter the address [http://xxx.xxx.xxx.xxx] and use the administrator account “admin” (or other preset accounts) to login. Click the [User manage system] button on the navigation bar of the platform.

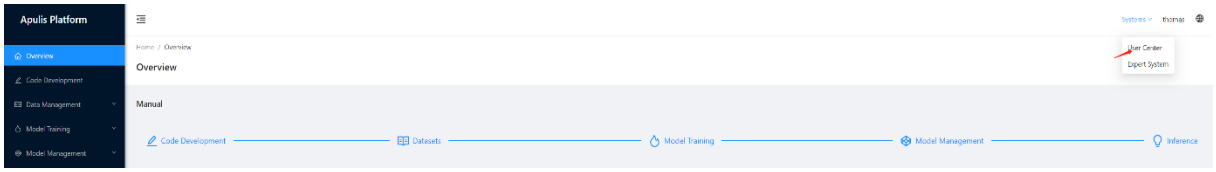


Figure 3: Enter user manage window

3.2.2 Create Ordinary User

Open the [User] menu on the left, you will see the initial administrators or user accounts of the system. The administrator can create a single user or multiple users by clicking [New User].

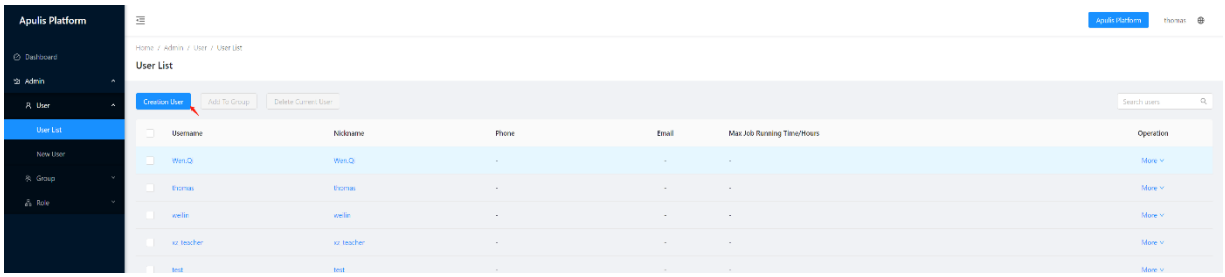


Figure 4: Create user account

Follow the instructions, fill in the required fields, and click next (you can leave the unrequired fields blank).

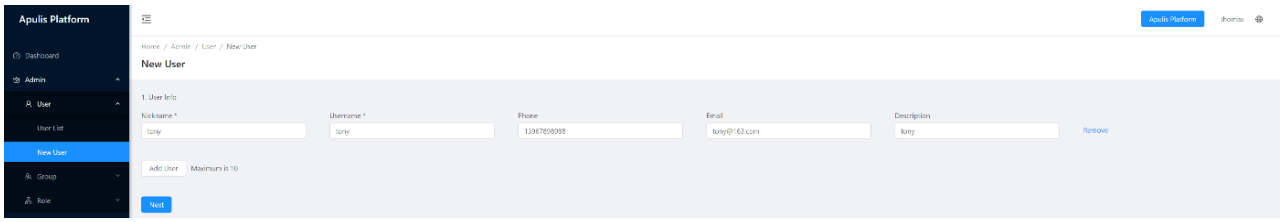


Figure 5: create user

3.2.3 Configure User Roles

Administrator can select roles for new users. The main roles include System admin, User, Annotation person. For general user, please select ‘User’

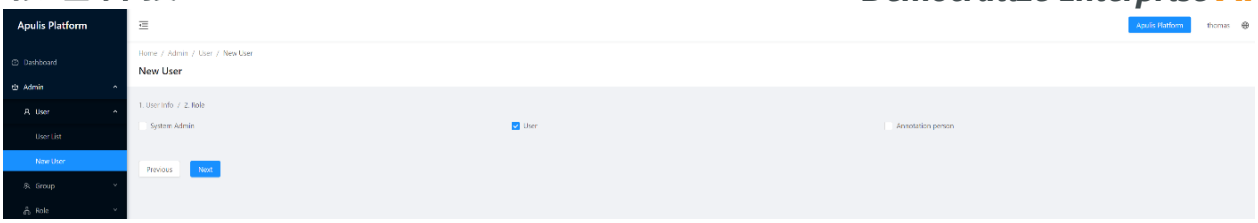


Figure 6: Select user roles

Click the [Next] button on Figure 6 and confirm the user information. By default, the system would initialize a complex password for new users. If you want to change the password, you can do it by clicking [Edit] on the right side.

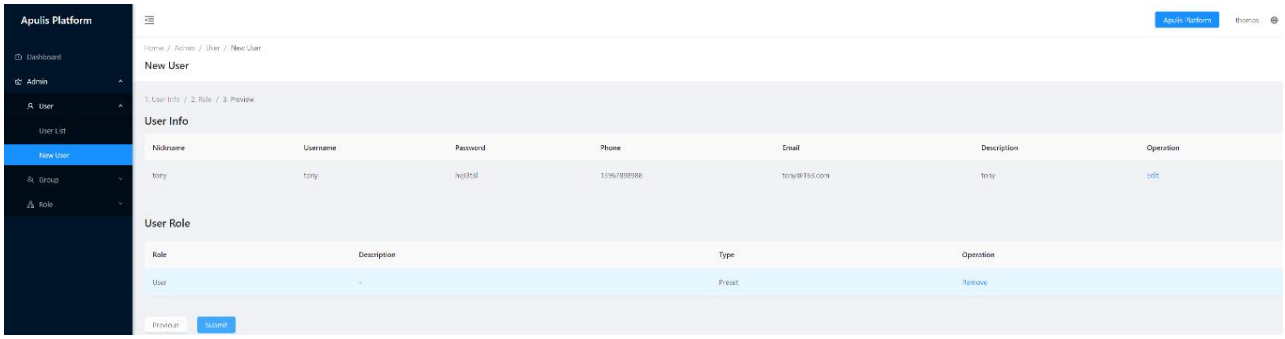


Figure 7: Confirm basic information of your account

3.2.4 Change Initial Password

After changing your password, click [Save] and then [Submit].

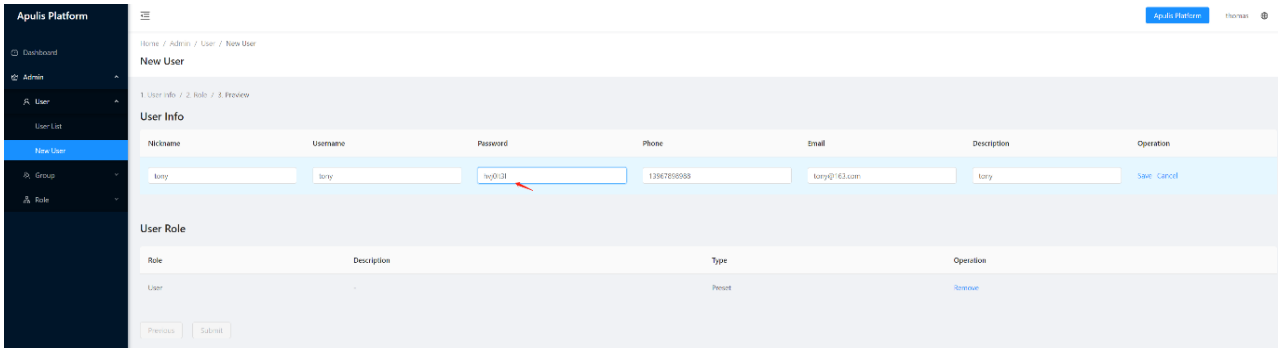


Figure 8: Change initial password

3.2.5 Associate Virtual Cluster

Return to the Apulis AI platform, click [Virtual Cluster] on the left menu bar and you will see your virtual clusters in a table. You can click [User] in the operation column to look at the existed users.

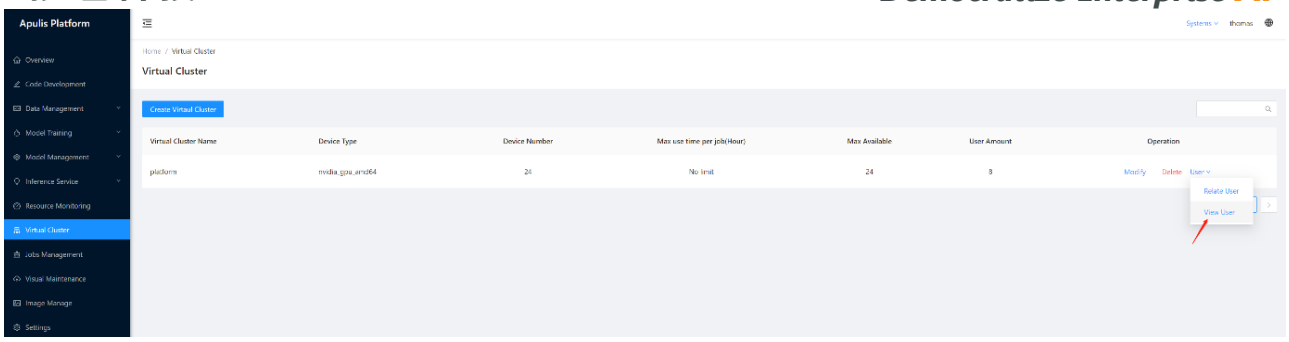


Figure 9-1: View the existed users on your default virtual cluster

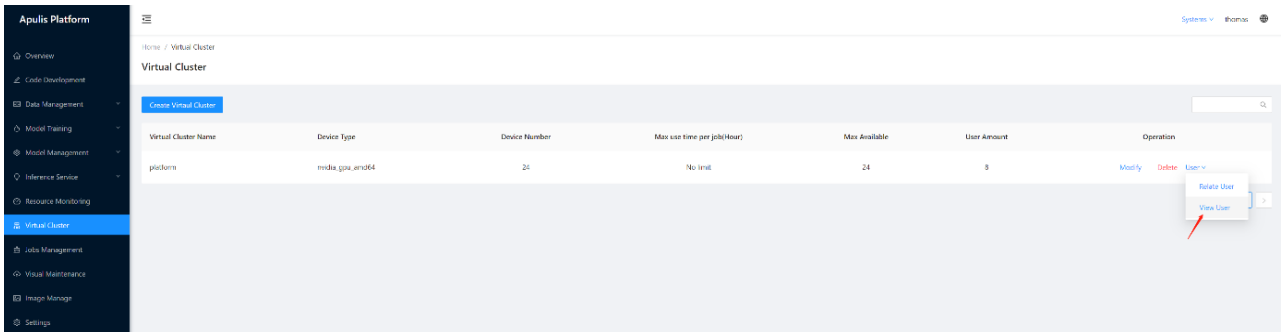


Figure 9-2: View the existed users on your default virtual cluster

Add users to your virtual cluster.

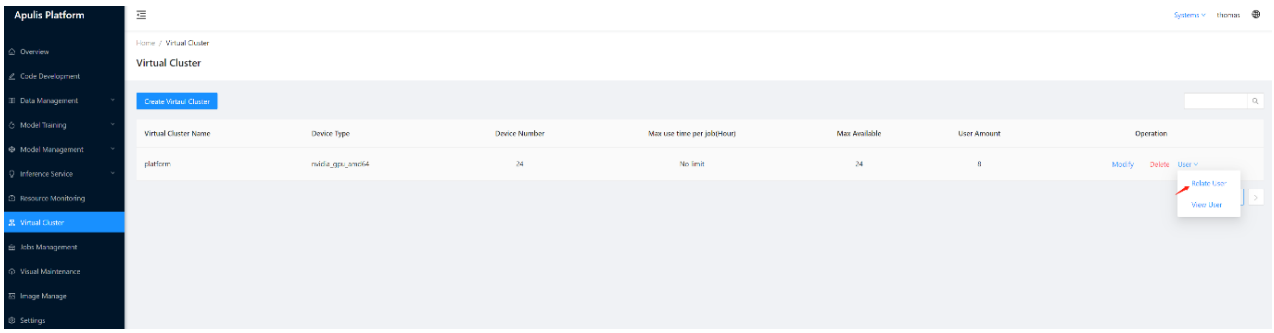


Figure 9-3: Add users to default virtual cluster

If the user has been successfully created, you can select the user and click OK, then the user is associated with the available resources.

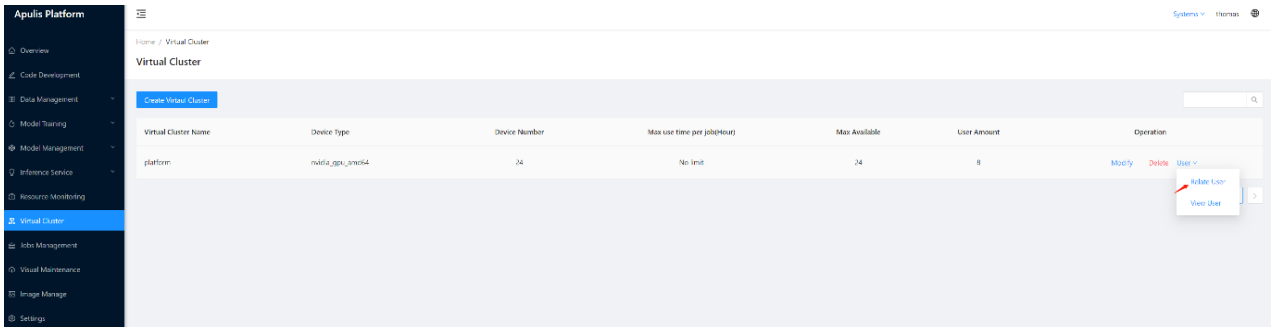


Figure 10: Associate users with virtual resources group

3.2.6 View Associated Virtual Clusters

After logged into the system, users can view the associated virtual clusters on the [Setting] menu. Users can switch to a different virtual cluster in the setting if they are associated with multiple virtual clusters.

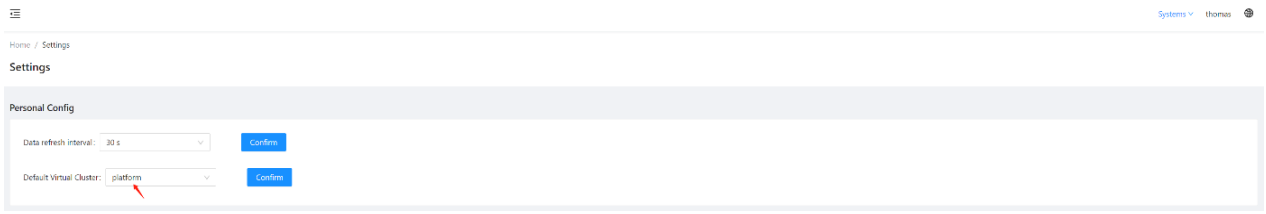


Figure 11: View associated virtual clusters

3.2.7 Manage Virtual Cluster

Only operations administrator can access the [Virtual Cluster] menu. In [Virtual Cluster] page, you can see a list of created virtual clusters.

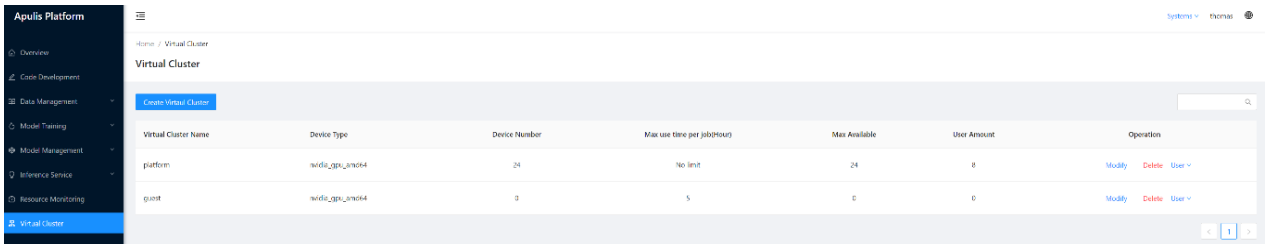


Figure 12: View Virtual Cluster list

Click [Create Virtual Cluster], the system will auto-detect the type of resource in clusters (such as huawei_npu_arm64) and maximum available resources. Administrators create a virtual cluster according to user's need and the remaining resources.

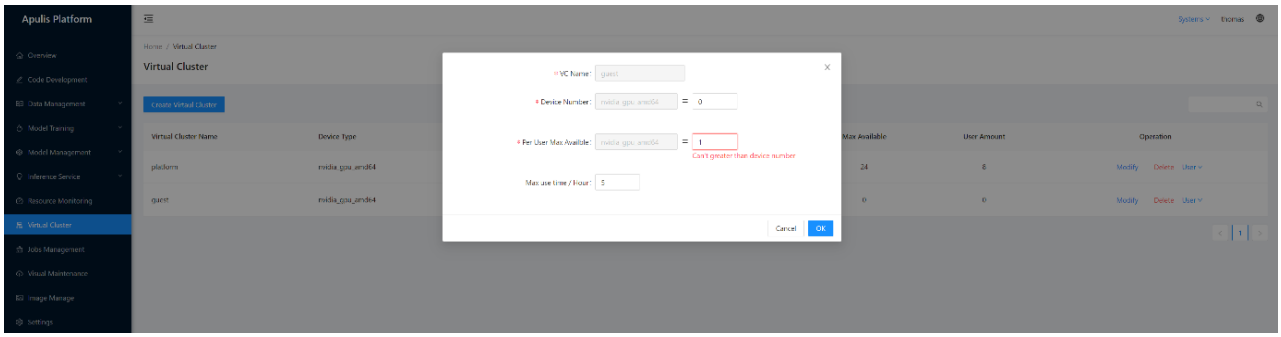


Figure 13-1: Create new virtual cluster – resource allocation

Administrators can set a time limit for virtual cluster to avoid resources being occupied for too long. Administrator can fill in the [Maximum usage time/hour] to set the time limit. Users can request for time extension after the time limit exceeded.

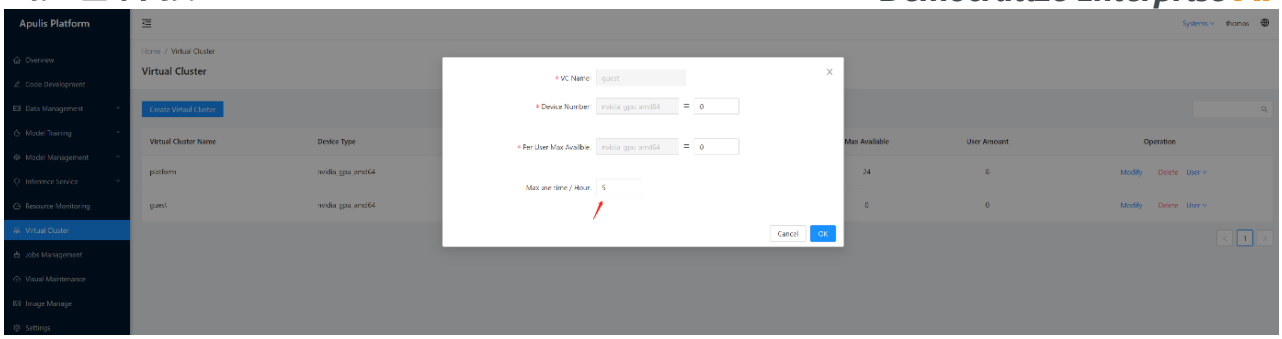


Figure 13-2: Create new virtual cluster – time limit

Administrator can update/delete the virtual clusters when no user is executing jobs on it.

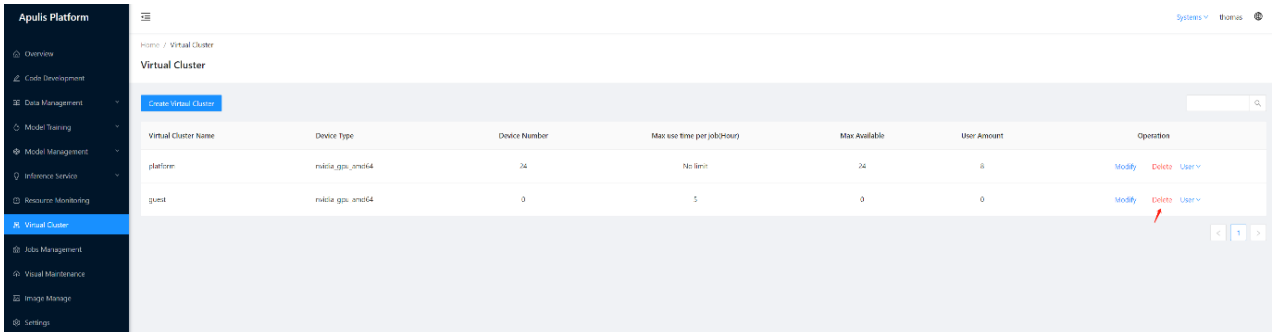


Figure 14: Delete virtual cluster

After administrator has added users into the virtual cluster, users will share all the resources in the cluster. When no resource is available, the submitted jobs will be queued up waiting for resources to be released. Administrator can add users to the virtual cluster or remove them.

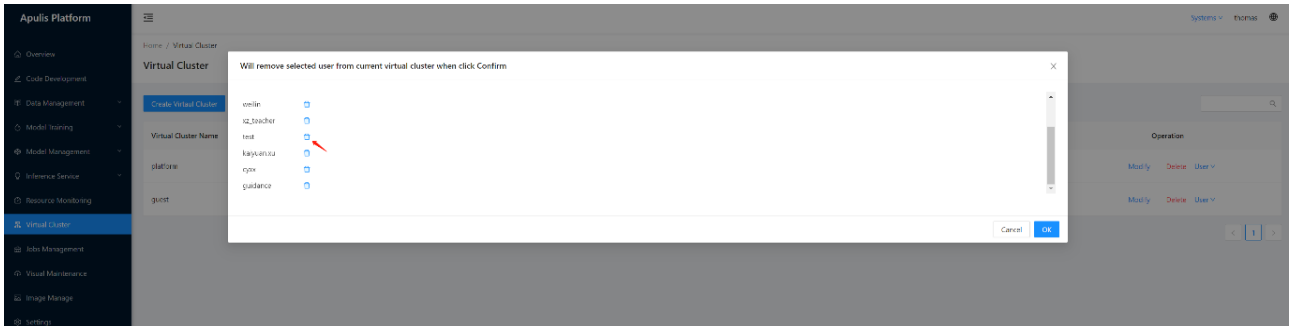


Figure 15: Add or remove users

3.2.8 Task resource limit

The administrator can pre-set the maximum number of resources available for the POD in each task on the NPU node; the default lower limit of the POD resource is CPU 2 cores, MEM 400 Mb; the upper limit is CPU 22 cores, MEM 80G; the administrator can be modify in the background according to the actual situation Configure the resource upper limit of the POD.

The configuration example:

1. Login the platform management node terminal and enter the installation directory
   `cd ~/InstallationYTung`
2. Modify the resource_limit item in the `group_vars/ cluster.yaml`:

```
resource_limit:
    huawei_npu_arm64:
      cpu:      22
      memory: 80Gi
```

3. Restart services `restfulapi2 and Jobmanager2`

   ./service_ctl.sh restart restfulapi2

   ./service_ctl.sh restart jobmanager2

4. Check the generated configuration

   cat /root/build/restfulapi2/config.yaml #

   resource_limit: {"huawei_npu_arm64": {"cpu": 22, "memory": "80Gi"}}

5. Check the task status in the resource monitoring window. The maximum resource usage in the POD has been limited to within 22 cores of the CPU and within 80G of memory.

FAQ:

1. The usage unit of the resource monitoring: CPU of the pod is based on the number of cores used, MEM 1G=1000kb; CPU of the node is based on the used time, MEM 1G=1000kb; the background monitoring is actually based on the number of CPU cores used, MEM 1Gi=1024kb .

2. When the MEM usage overrun the limit, such as a memory leak or deadlock in the training process, the platform will stop the task and the status in the task list will be displayed as "failed"

3.3 Overview

The navigation bar at the top of the page consists of two parts: system navigation drop-down menu (includes link to expert system and user management system) and the currently logged in user.

The left menu bar includes 12 items: Overview, Code Development, Data Management (Datasets), Model Training, Model Management, Inference Service, Resource Monitoring, Virtual Cluster, Task Management, Visual Operation and Maintenance, Image management and Setting. Data Management includes dataset management and data annotation. Model Training includes model training and training parameter management. The Model Management includes my models, preset models, evaluation list and evaluation parameter management. The Inference Service includes central inference and edge inference.

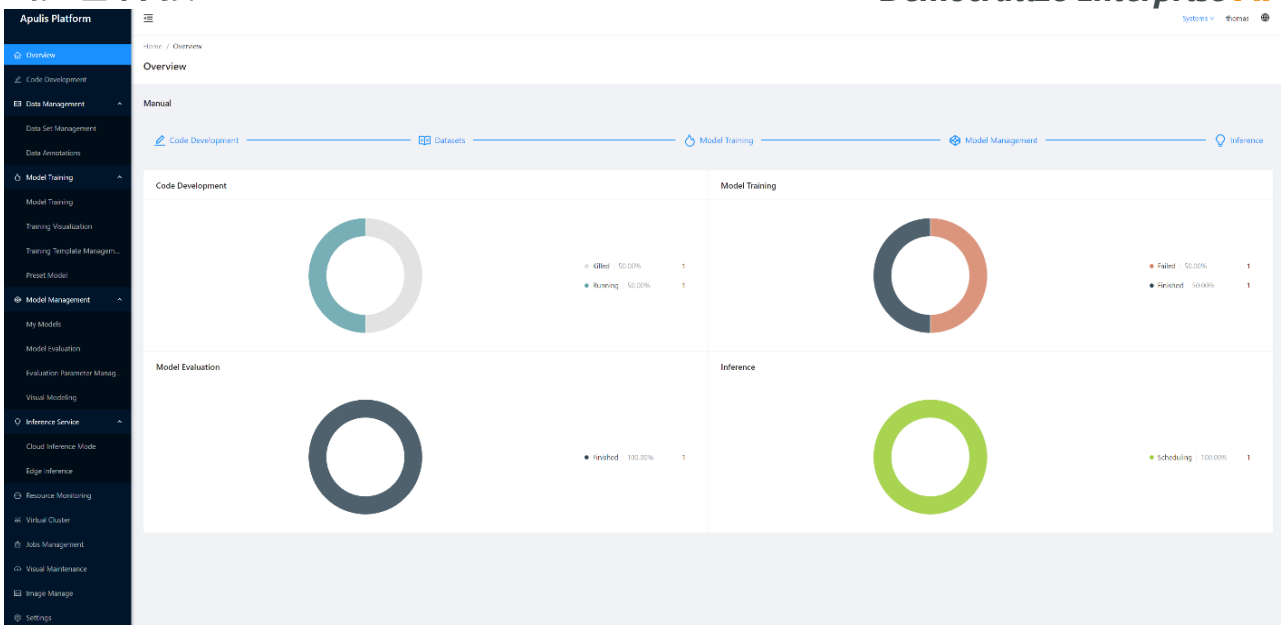


Figure 16-1: Overview

[Virtual Cluster] menu and [Jobs Management] menu are only visible to administrators. Ordinary users do not have access to these menus.

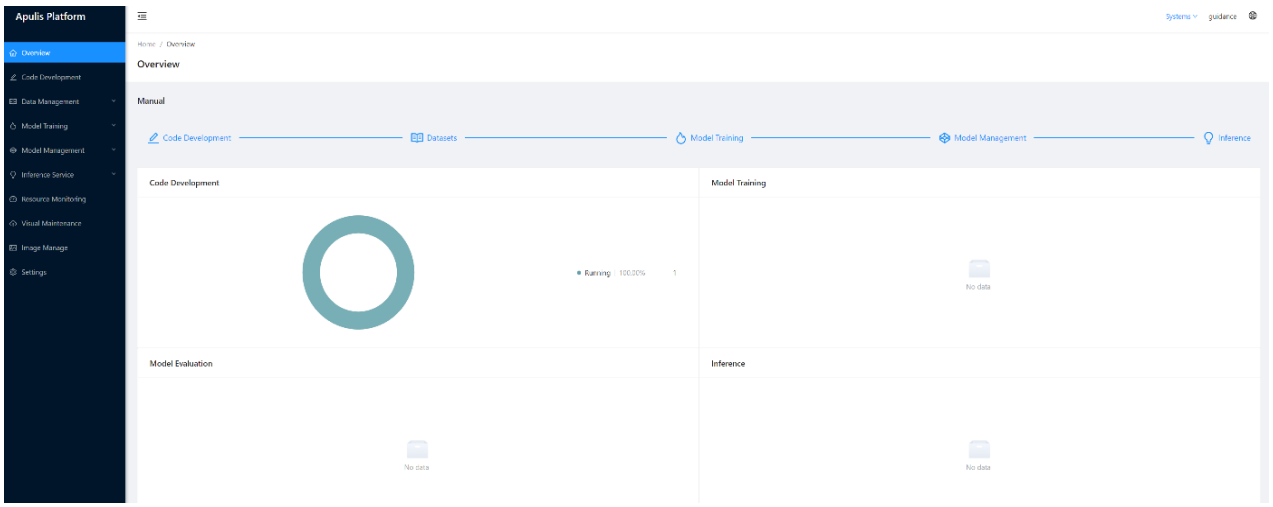


Figure 16-2: Normal user account overview

3.3.1 Menu Bar

The navigation bar at the top of the page is shown in Figure 17. You can click the left-most icon  to show or hide the left menu bar. Click "Apulis AI Platform" to jump to the Home Page. Click the menu "User Management System" to jump to the user management page, and the current logged-in user is displayed on the far right.

Figure 17: Top Navigation bar

3.3.2 Development Pipeline

The Development pipeline is displayed at the top of Overview page.

We recommend users to develop their AI applications in the following order:

- Code development -- create a code development environment for users to execute their modeling scripts.
- Data Management (Datasets) -- upload datasets and annotate data if necessary.
- Model Training -- load the annotated dataset and create model training jobs.
- Model Management – manage the trained models.
- Inference – central side interference service.

Figure 18: Development pipeline

3.3.3 Statistic Charts

The charts illustrated the percentages of jobs in different status in code development, model training, model evaluation and inference services, so that users can clearly understand the current task situation.

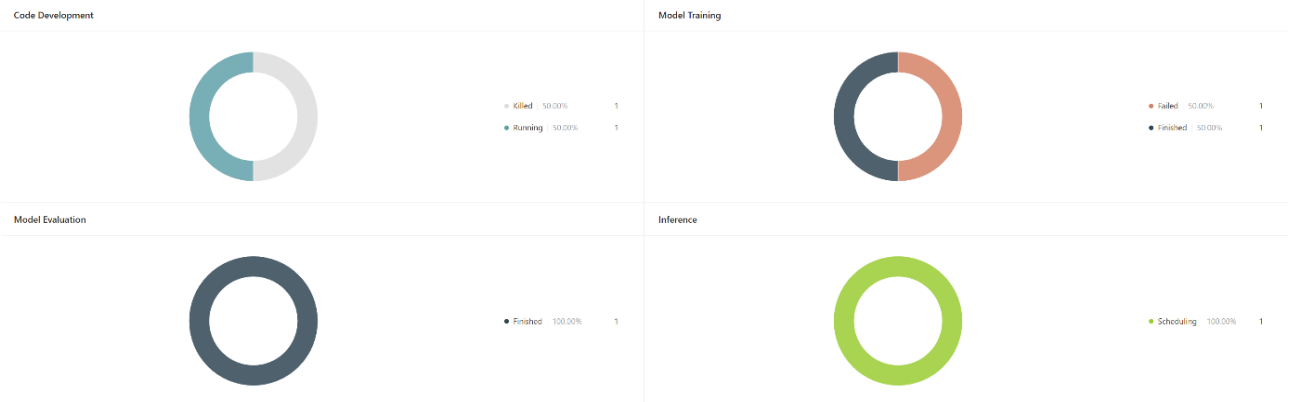


Figure 19: Statistic chart

3.4 Code development

3.4.1 Create Development Environment

Click "Code Development" in the menu bar and then click "Create Development Environment" button to enter the Create Development Environment page, as shown in Figure 20. After submitting the request, it jumps to the code development environment list page.

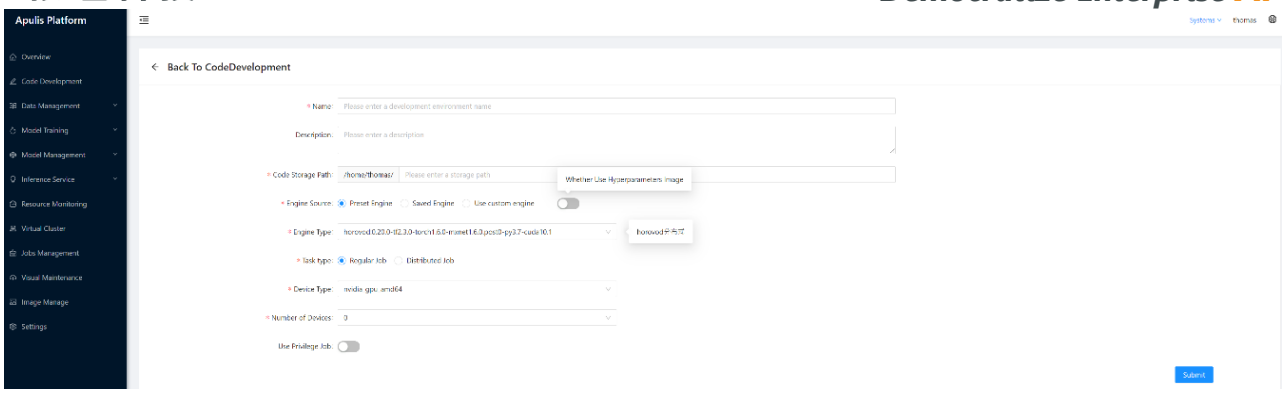


Figure 20: Create new development environment

The basic environment settings include the development environment's name, description, code storage path, engine type, task type, device type, and device number. Explanations of these settings are as follows:

- Development environment's name: required field, name for your development environment.
- Description: optional field, description for user-defined development environment.
- Code store path: required field, the directory to store code files. User's working directory would be `/hone/{username}`.
- Engine type: required field, machine learning algorithm framework.
- Task type: required field, type of the task, either [a regular task] that runs on a single server or [a distributed task] that runs on multiple servers.
- Device type: required field, type of processor. Currently, our platform supports Huawei_NPU_ARM64 and Nvidia_GPU_AMD64.
- Number of devices: required field, the number of devices. Only 0, 1, 2, 4 and 8 is valid as specified by the manufacturer.
- Number of nodes: required field, it is only necessary for [Distributed task]. it specifies number of physical nodes required for training.
- Number of devices on a single node: required field, it is only necessary for [Distributed task]. it specifies the number of devices on each node.
- Total number of devices: this field shows the number of total devices requested for distributed tasks. Total number of devices = number of nodes × number of single node devices.

3.4.2 Code Development Environment List

The [Code Development Environment List] displays a list of all code development environments created by the current user as shown in figure 21.

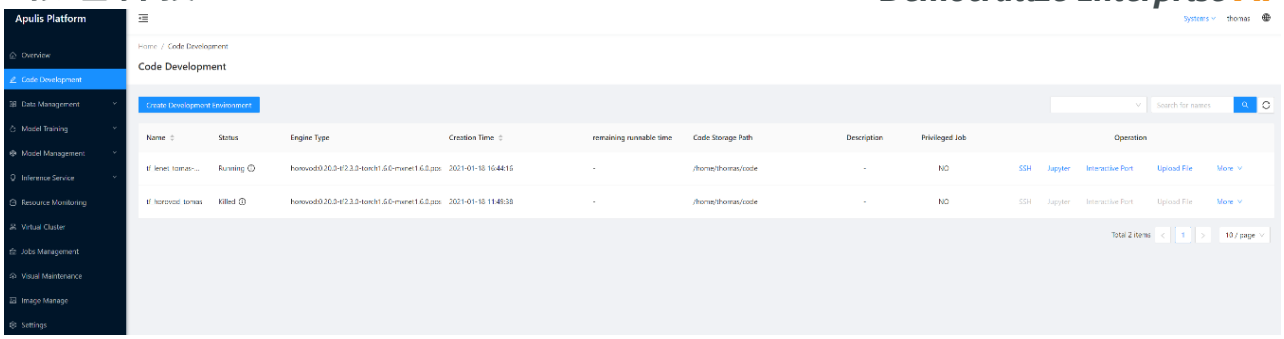


Figure 21: Code development environment page

The table includes 7 columns: development environment name, state, engine type, creation time, code store directory, description, and operation.

- Development environment name: Name entered by the user at creation time.
- Status: the running state of the environment, including unapproved, queued, scheduled, running, stopped, closed, error, etc.
- Creation time: Time when user submits the job creation.
- Code store directory: The path that the user filled in at job creation time.
- Description: description filled in by the user at job creation time.
- Operations: You can open, upload, and stop a running development environment.

3.4.3 Jupyter Environment

To start a Jupyter Environment, please choose a running environment and click [Jupyter] in the operation column. In Jupyter environment, users can develop code.

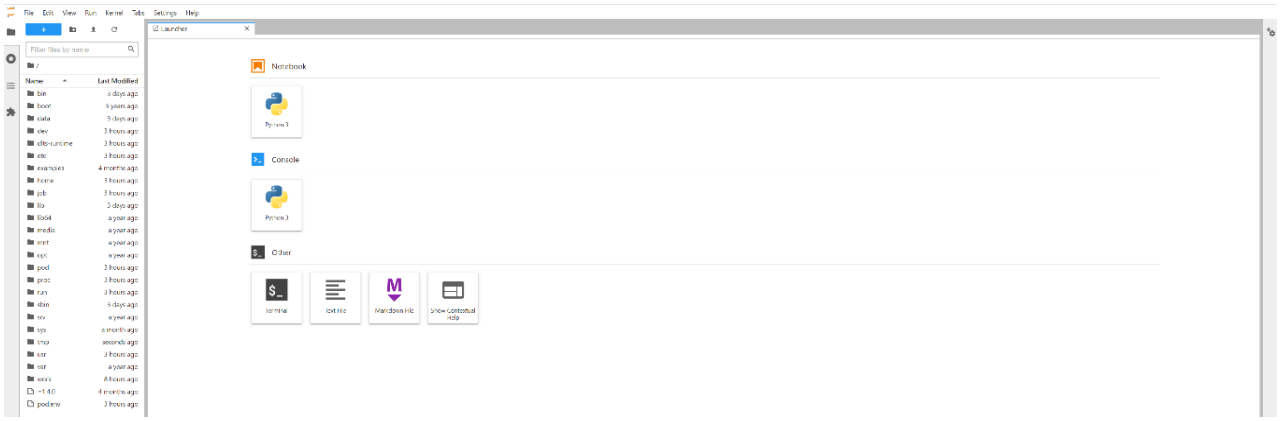


Figure 22: Jupyter Environment

3.4.4 Upload Code

Users can also code in their local IDEs, and then upload their code into the development environment by clicking [Upload Code] in the operation column of a running environment. The uploaded files will be stored in a code directory. Note that files exceed 500MB is recommended to upload through SCP.

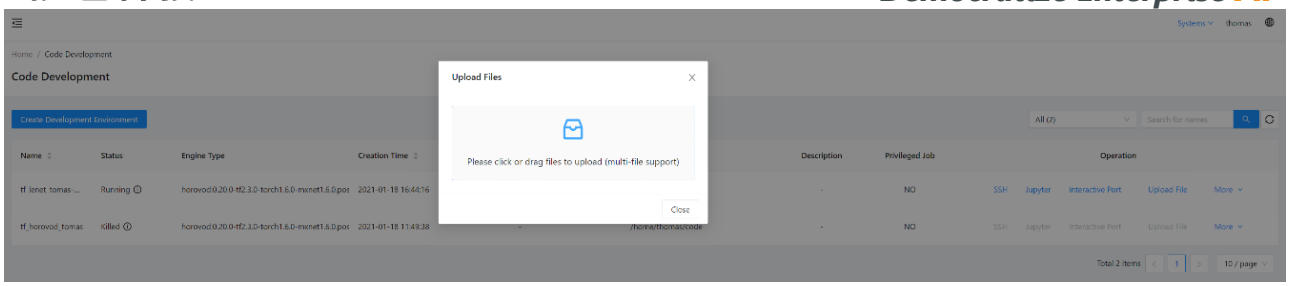


Figure 23: Upload code modal

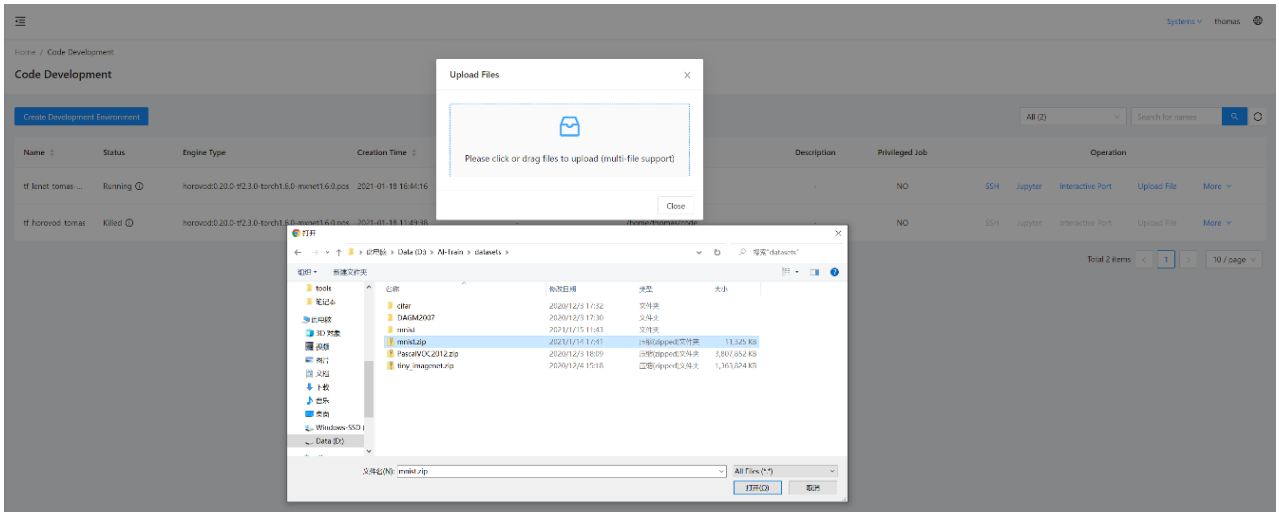


Figure 24: upload code – select files modal

3.4.5 Acquire SSH link

Select a running environment and click the [SSH] link in the operation column, the page would display the information needed to access the code environment through SSH.

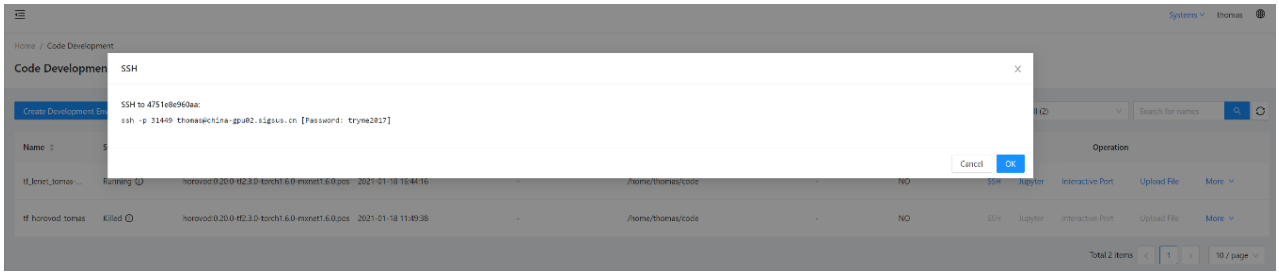


Figure 25-1: Acquire SSH link

When the environment is running, users can access the development environment through SSH link. Note that: the default link contains the certificate of the cluster, please use the following IP address to access the server:

IP: ssh -p <PORT> <USERNAME>@<DOMAINNAME OR IP>

Password: tryme2017

Figure 25-2: Login through SSH

3.4.6 Stop Development Environment

Users can stop a running environment by clicking the [more->stop] in the operation column. The stopped environment would change its state from [Running] to [Closing], and eventually [Closed]. After the environment is closed, all operation buttons would be disabled. Files in the environment would not be removed (stored in [Code store path]) after the environment is closed. Closed environment cannot be re-open again, but you can recreate the environment with the same configuration.

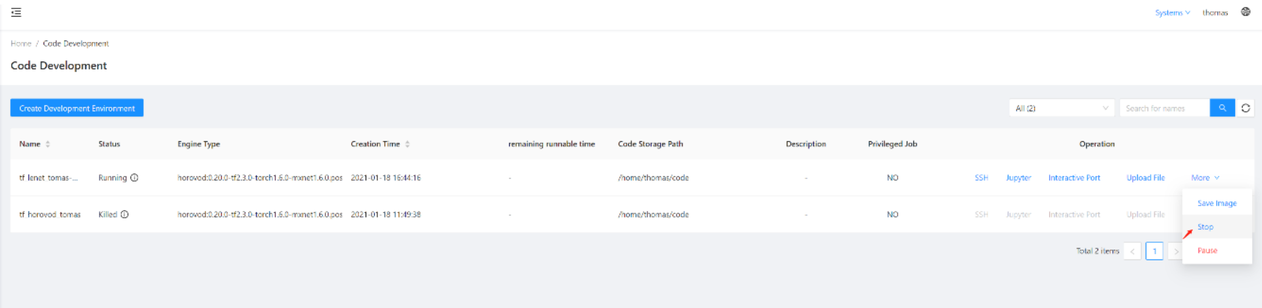


Figure 26: Stop development environment

3.4.7 Delete Development Environment or Save Image

When a closed environment is no longer needed, users can remove the environment by clicking [more->delete] in the operation column.

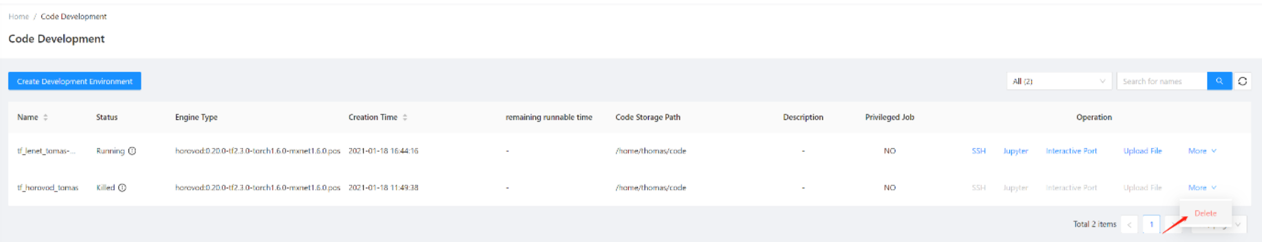


Figure 27-1: delete development environment

If you want to save the images in your environment, you can do it by clicking [more->save] in the operation column of a running environment.

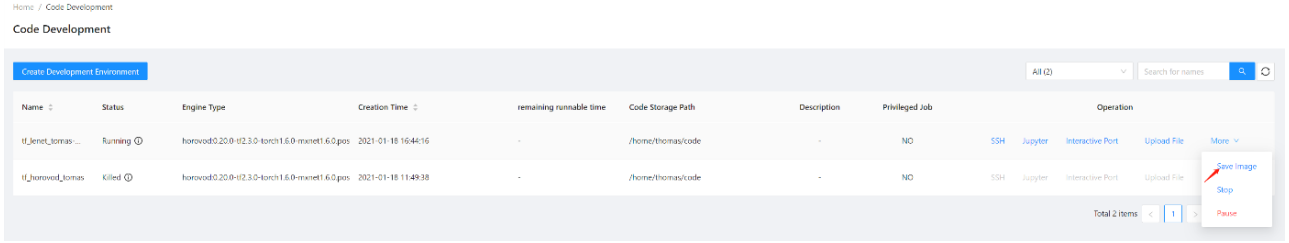


Figure 27-2: save images

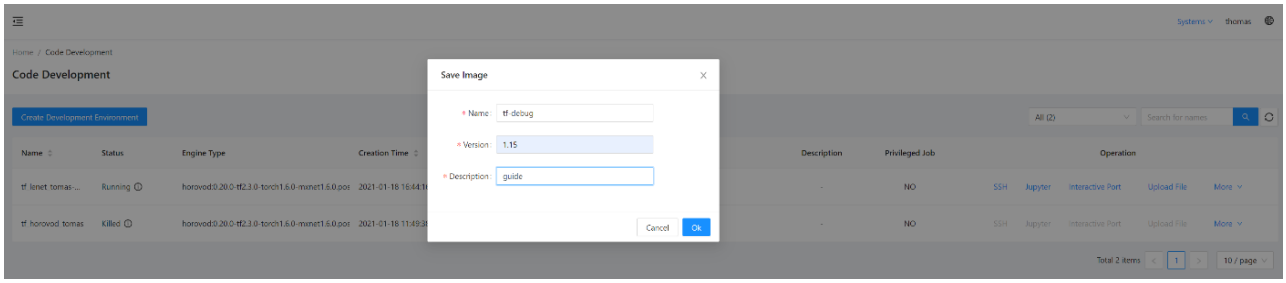


Figure 27-3: configurations for image saving

3.5 Data management

[Data management] module can help users edit uploaded datasets, annotate datasets or transform format of datasets.

3.5.1 Create new dataset

By Clicking [Data management-> Dataset Management] in the menu bar and then the [Add dataset] button, the add new dataset modal window would pop up, as shown in figure 28-1 (upload through browser) and figure 28-2 (upload through other methods).

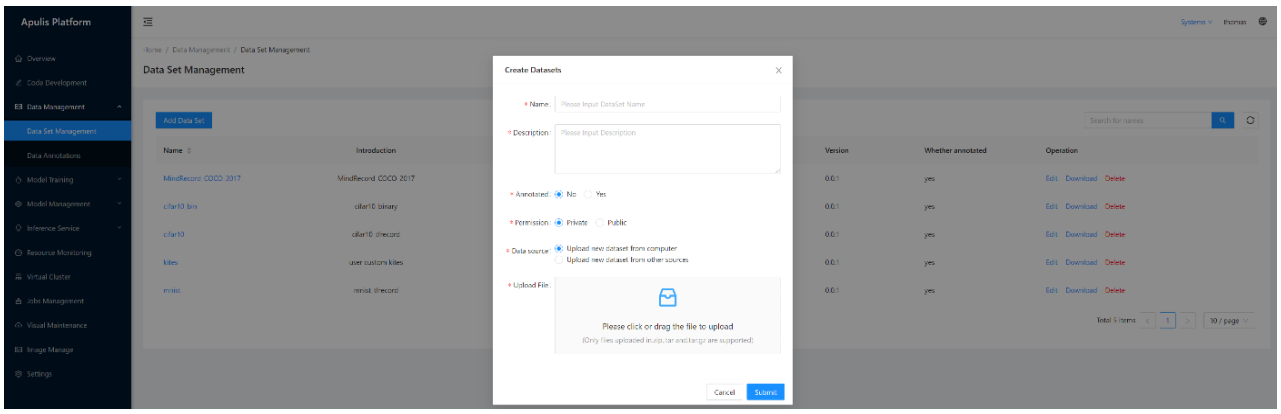


Figure 28-1: Add new Dataset modal

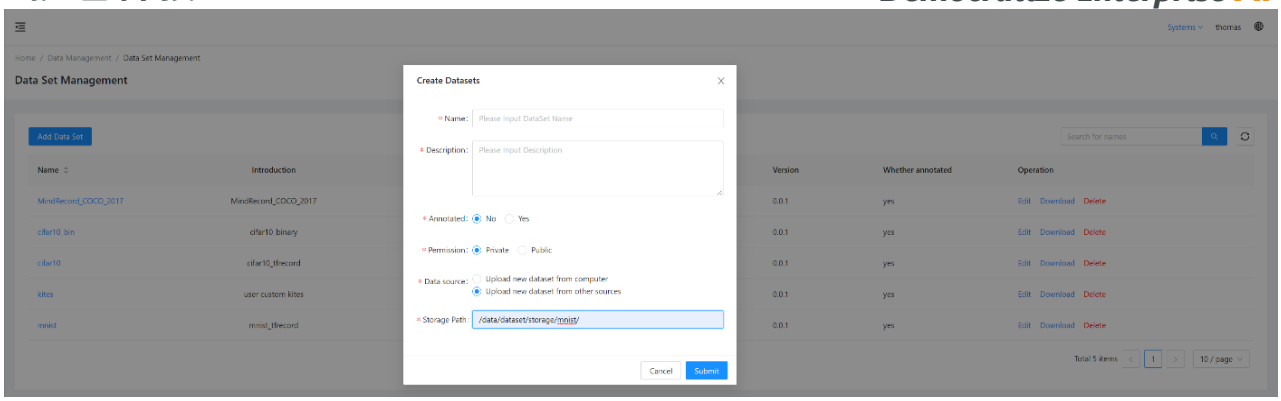


Figure 28-2: Upload dataset through other methods

The form of adding dataset includes 7 fields, dataset's name, brief introduction, whether it is annotated, data permission, data source, upload file, storage path.

- Data set name: required field, user can customize input.
- Description: required field, user's description of the dataset.
- Whether annotated: required field, 'Yes' means the uploaded dataset is already annotated and can be directly used in training, while 'No' means it is not annotated and should be annotated before training.
- Data permissions: required field. 'Private' means the dataset is only visible to the currently logged in user, while 'Public' means it is visible to all users in the system.
- Data source: required field, the dataset uploaded through browser or other sources.
- Upload file: Only required when dataset is uploaded through browser. Users can select compressed dataset files from their local storage. Note that only '.zip', '.tar', '.tar.gz' files are supported and the dataset file should not exceed 2GB (files larger than 2GB is recommend to upload through other sources). Uploaded datasets would be decompressed automatically.
- Storage path: Only required when dataset is uploaded through other sources. If your dataset is larger than 2GB, it is recommended to create the dataset this way. Users can upload their dataset to the server, decompress it manually, and then fill in this field with the dataset's absolute path in the server.

3.5.2 Dataset Management List

This page shows a list of datasets that the current user created or can operate on.

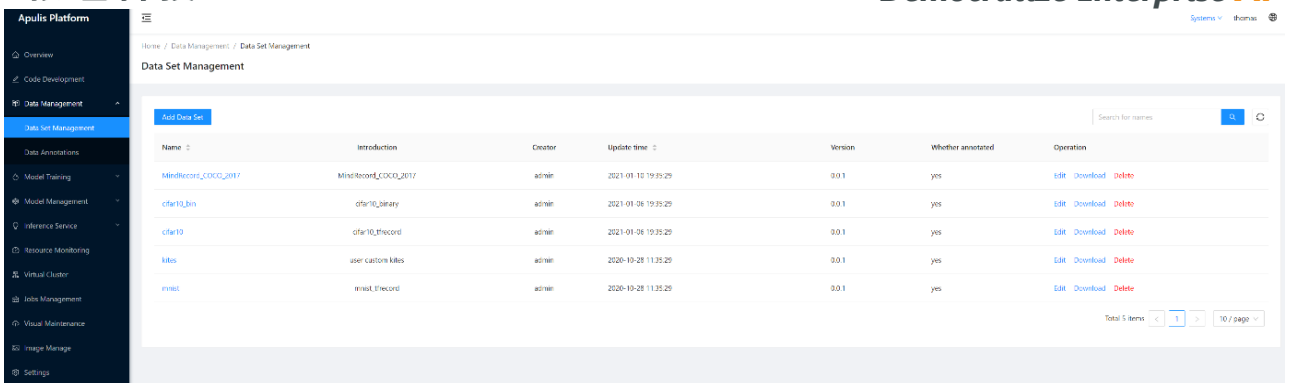


Figure 29: Dataset list

The dataset list includes 7 columns: dataset's name, description, creator, update time, update version, whether annotated, operation.

- Dataset name: The name that the user filled in at creation time.
- description: The description filled in by the user at creation time.
- Creator: Creator of the dataset.
- Update time: latest date that the dataset was updated.
- Updated version: latest version of the current dataset.
- Whether annotated: whether the dataset is annotated. Only annotated dataset can be used in training.
- Operations: Operations that can be applied to the dataset, including editing, downloading, deleting.
- Datasets in the list can be edited, downloaded, deleted. The dataset edit modal is shown in figure 30.

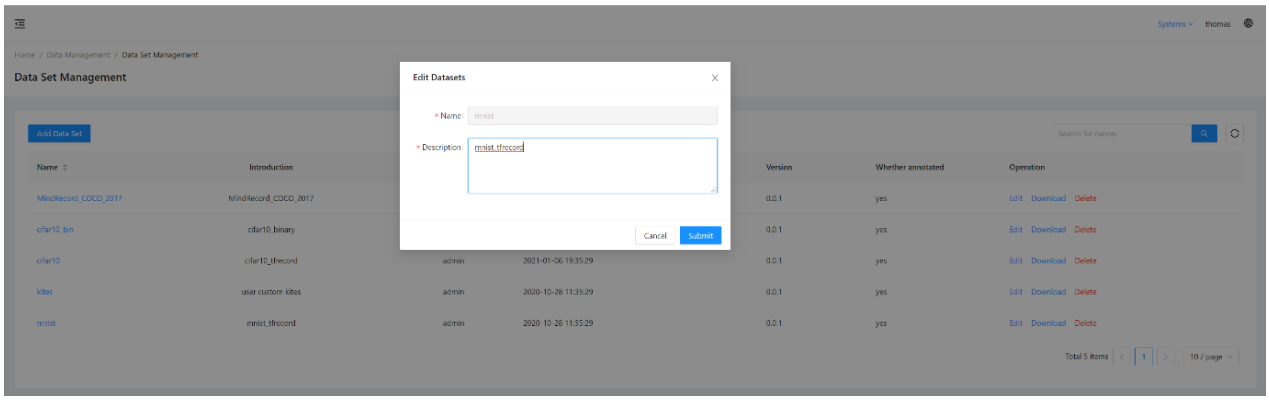


Figure 30: Edit dataset

3.5.3 View Dataset Details

You can enter the [Dataset Detail] page by clicking on the dataset's name.

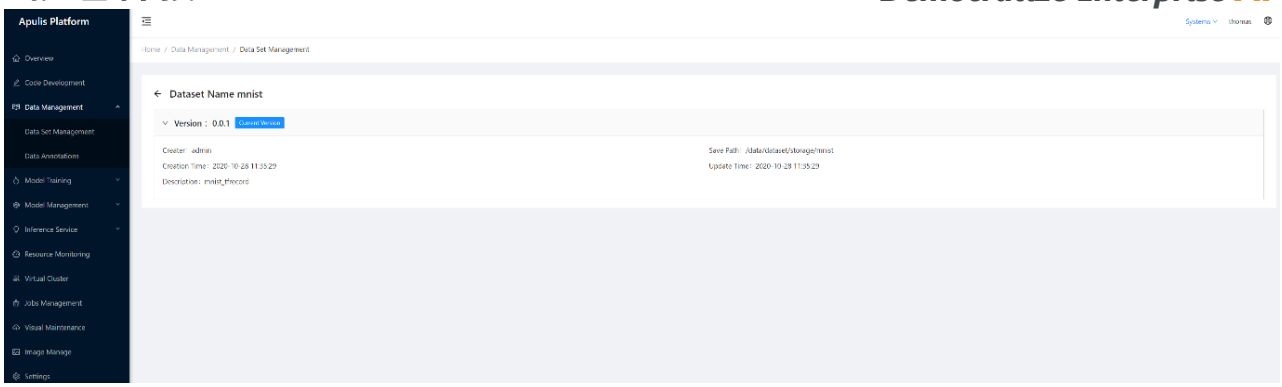


Figure 31: Dataset detail

The dataset detail page shows 7 field: data set name, version, creator, creation time, description, storage path, update time.

3.5.4 Data Annotation Platform

You can enter the data annotation platform by clicking [Data Management -> Data Annotation] in the menu bar.

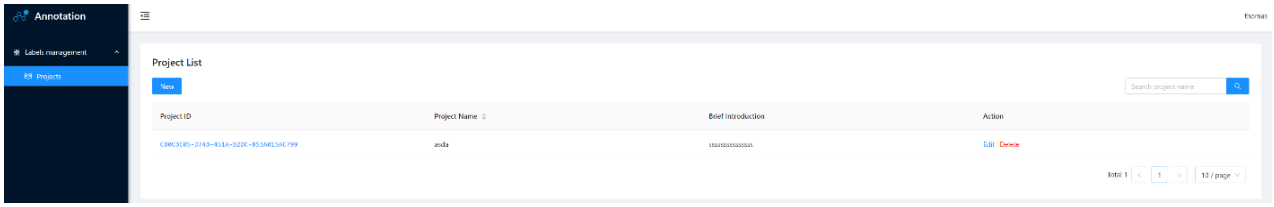


Figure 32: Data Annotation Platform

You can manage your annotation project in this platform. One task can contain multiple datasets.

3.5.5 Create News Annotation Project

Click [New Project] button, fill in the fields in the [new project modal]. There are two fields: the name of the project and the description.

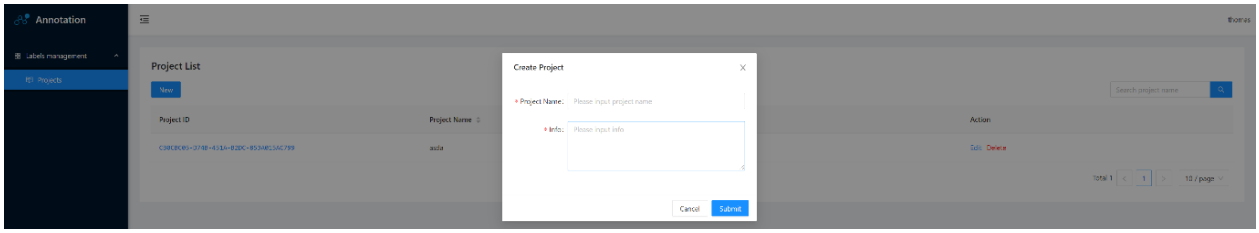


Figure 33: New Project Modal

3.5.6 Annotation Project List

The list of projects that created by the current user.

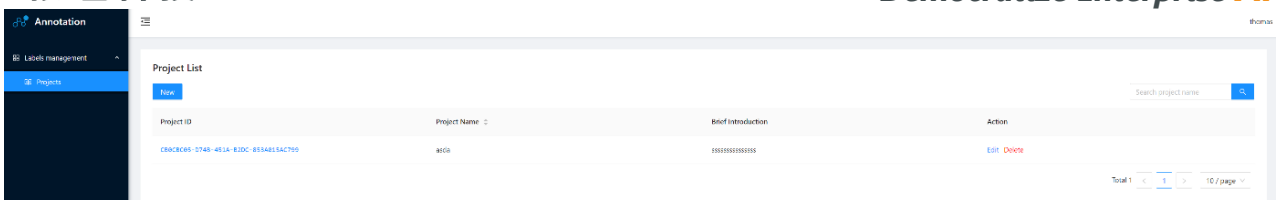


Figure 34: Project list

The annotation project list includes four columns: project ID, project name, a brief description and operations. The possible operations can be either: 1) to edit (Figure 35 and, 2) to delete the specific annotation project.

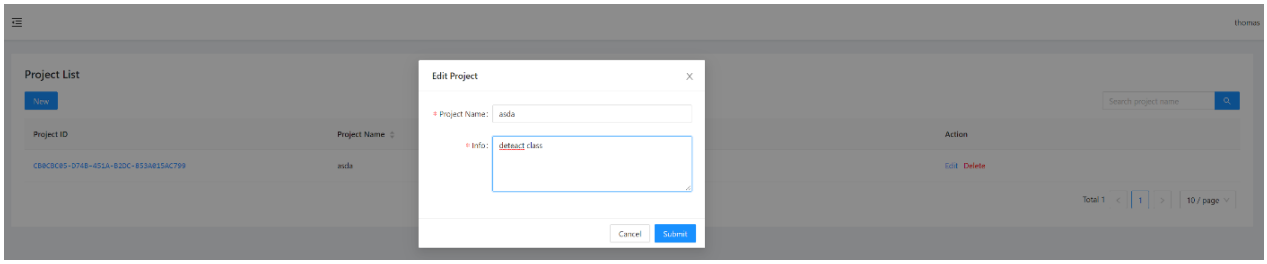


Figure 35: Edit annotation project

3.5.7 Dataset List in Annotation Project

Click the [Project id] of a project, and you will enter the dataset list page of this project.

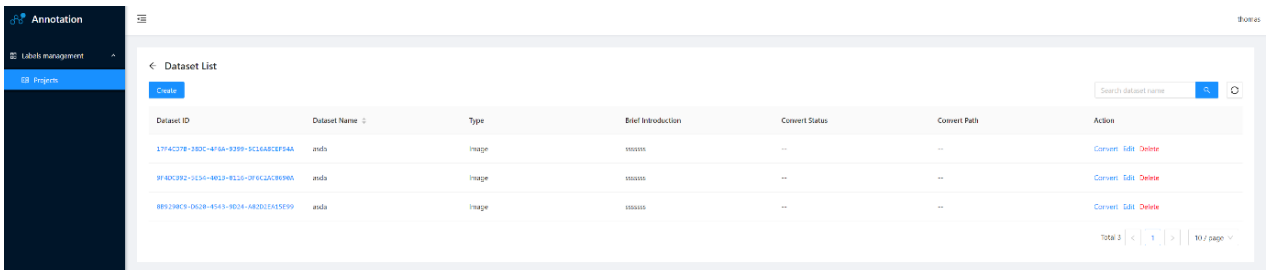


Figure 36: Dataset list in annotation project

3.5.8 Create New Dataset

Click [add dataset] and fill in the fields in the pop-up modal to create a new dataset.

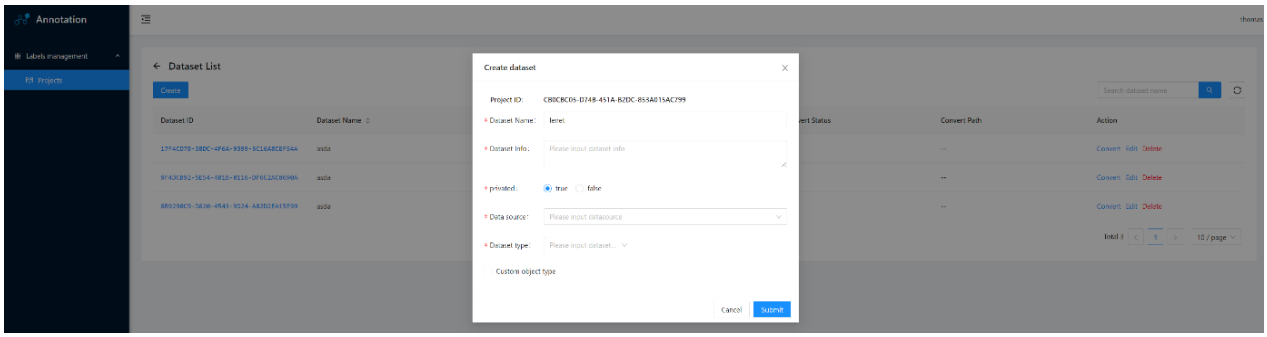


Figure 37-1: create new dataset pop-up

There are 7 fields in the new dataset form: project ID, dataset name, a brief description, permissions, data source, dataset type, and custom object type. The detailed descriptions are as follows:

- Dataset name: required field, name of the dataset.
- A brief description: required field, description of the dataset.
- Permissions: required field, Either 1) "Private", which it is only visible to the current user, or 2) "Public", which means it is visible to all users in the system.
- Data source: required field, the data source that needs to be annotated. You can only select data sources that are not annotated.
- Dataset type: required field. Currently, only image dataset type is supported. The platform will add other dataset types in the future.
- Custom object types: optional field. Currently, we support polygon object type and bounding box (bbox) object type. A polygon is an arbitrary polygon annotation and a bbox is a rectangular annotation.

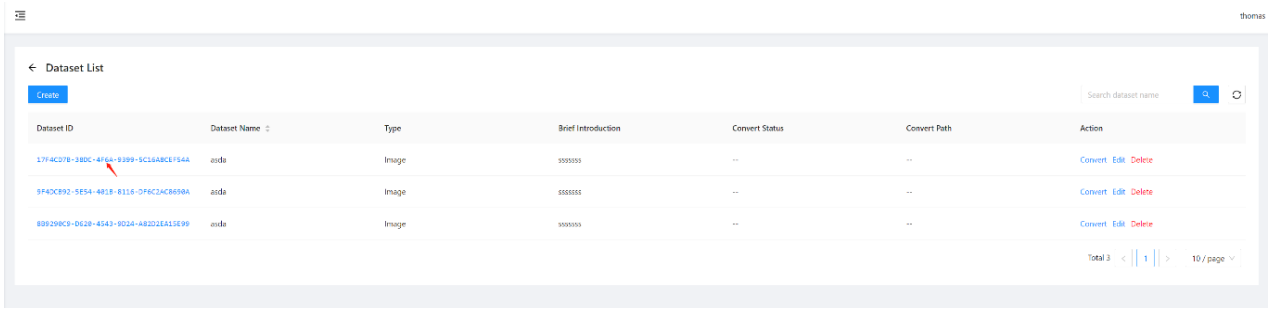


Figure 37-2: Annotation page

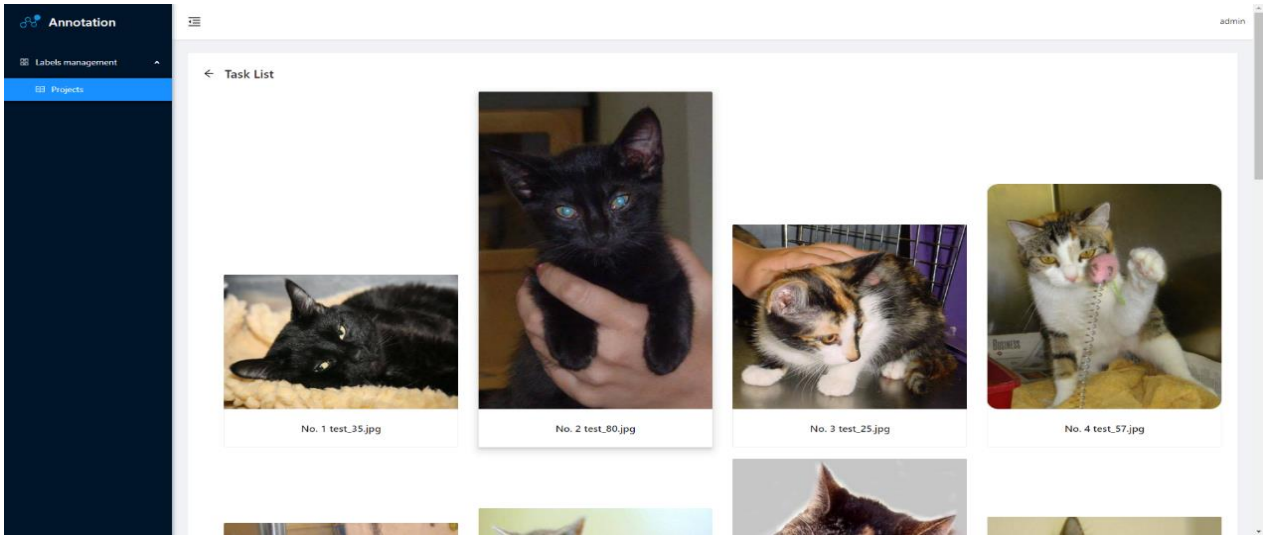


Figure 38: Task List

3.5.9 Image Annotation

Click an image in the task list to annotate it. Select object type and then click the plus buttons to create an annotation; You can delete a object type or a annotation by clicking the delete button ( ). Clicking the view button ( ) can hide or show an annotation

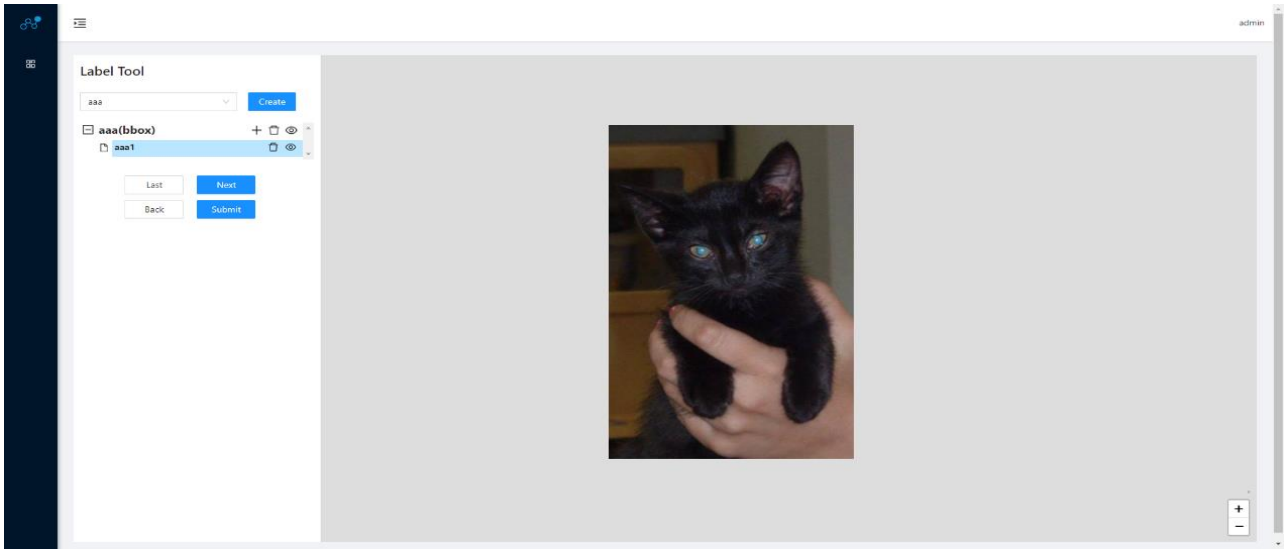


Figure 39: Annotation page

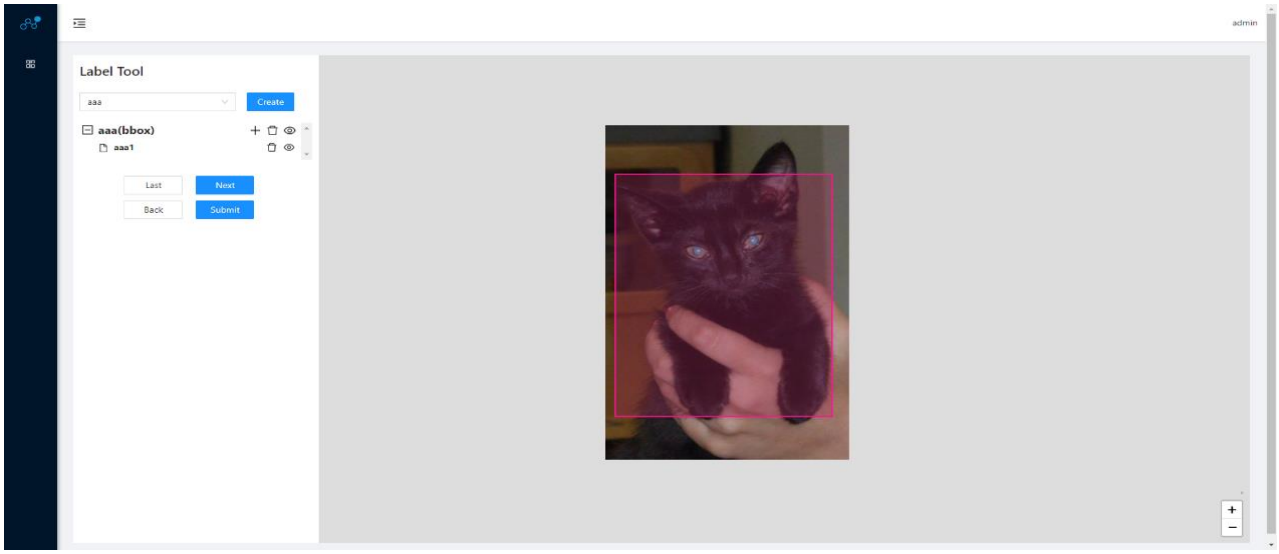


Figure 40: A finished example

Your annotation won't be saved until you click [Submit]. After you click [Submit], it will switch to next image until all images in the dataset are finished.

3.5.10 Dataset Format Transformation

In the [Dataset List] page, you can click the [Convert] button to transform the dataset's format. Currently, our system supports conversion to COCO dataset format.

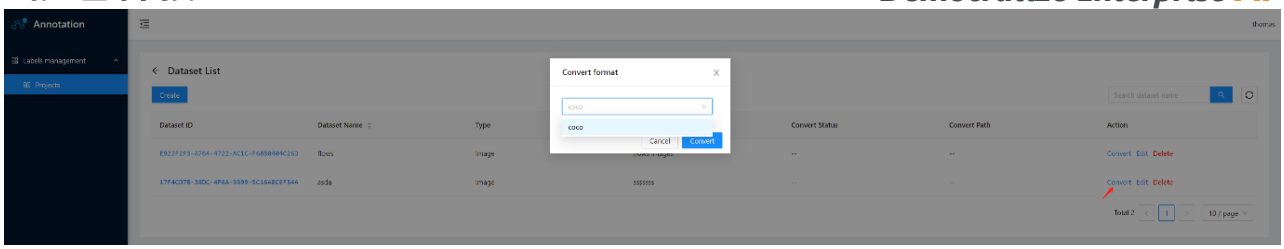


Figure 41: Dataset format transformation

3.6 Model Training

This module includes create model training job, training parameters manage and preset model list.

3.6.1 Model Training Job

The [Model Training Page] shows a list of trainning jobs, which shows the name, status, type and create time of the trainning job.

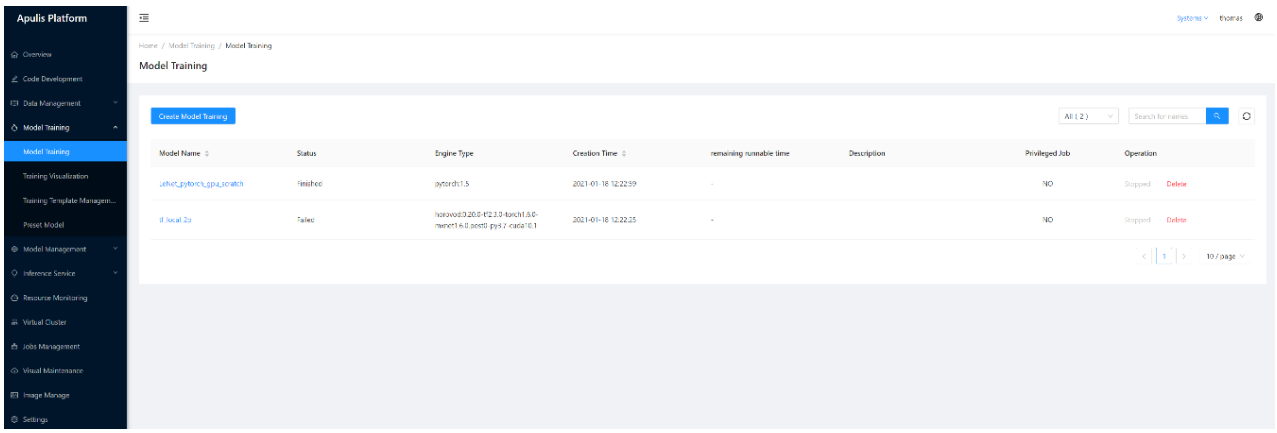


Figure 42: Model Training Job List

Click the [Create Model Training] button and fill in the required filed to create a training job. Please add dataset name into the parameter configuration.

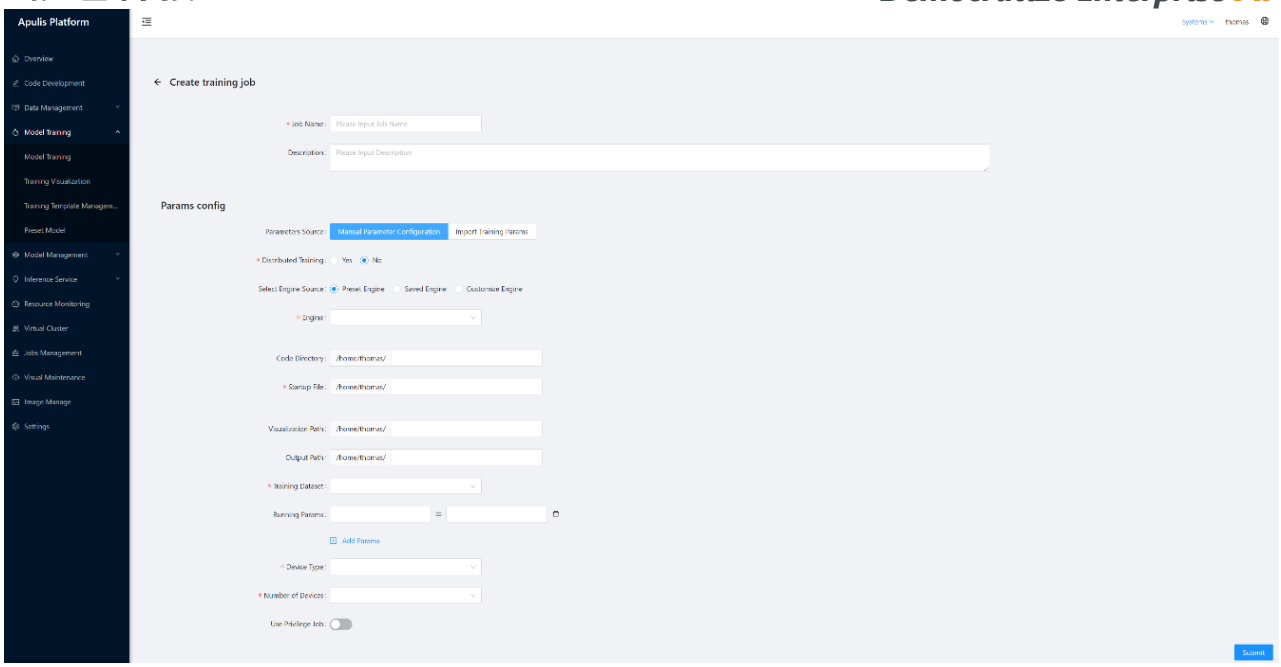


Figure 43: Create model training job

You can stop the training by clicking [stop] in the operation column.

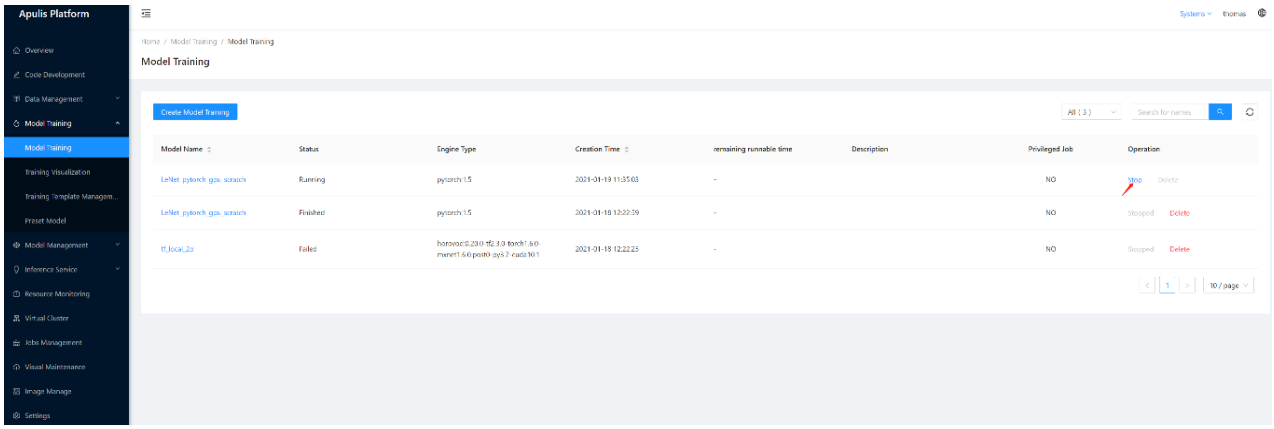


Figure 44: Stop training

3.6.2 Training Template Management

To make it easier for you create a training job, you can save you training settings as a template so that you can create a new job from it. We also provided the preset model's template.

1. Template list

   In the [Training template management] page, there are a list of templates.

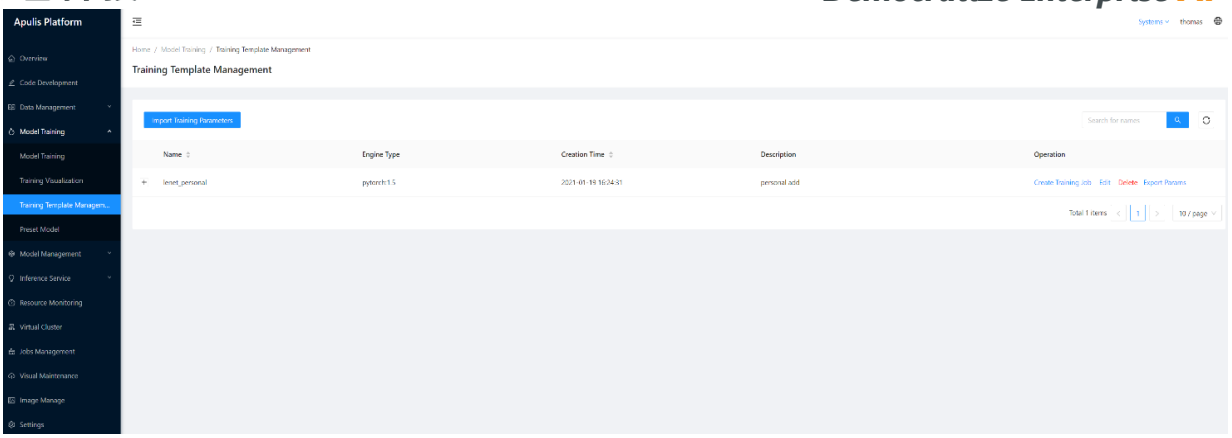


Figure 45: training template list

2. Update preset template

   Click [Edit] in the operation column to update the template.

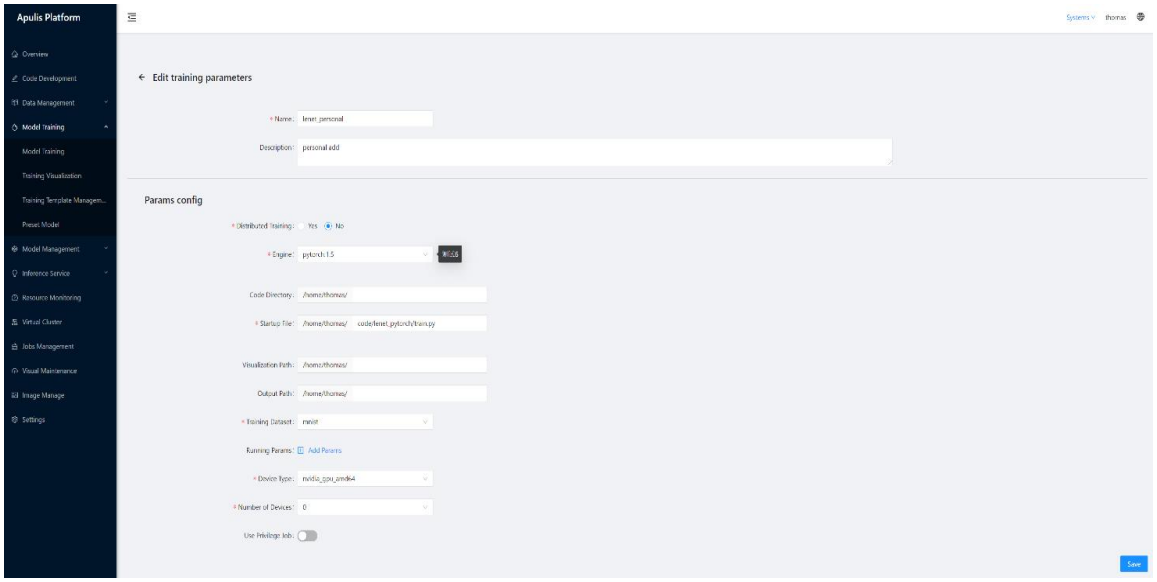


Figure 46: Edit or update preset training template

3. Create training job from template

   Click [Create Training Job] in the operation column to create a new training job from the template.

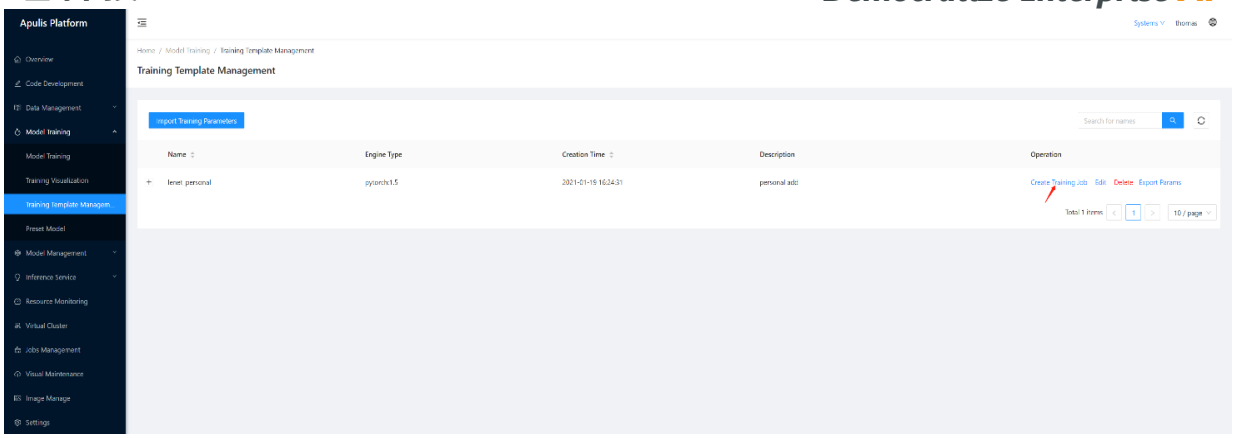


Figure 47-1 Create training job from template

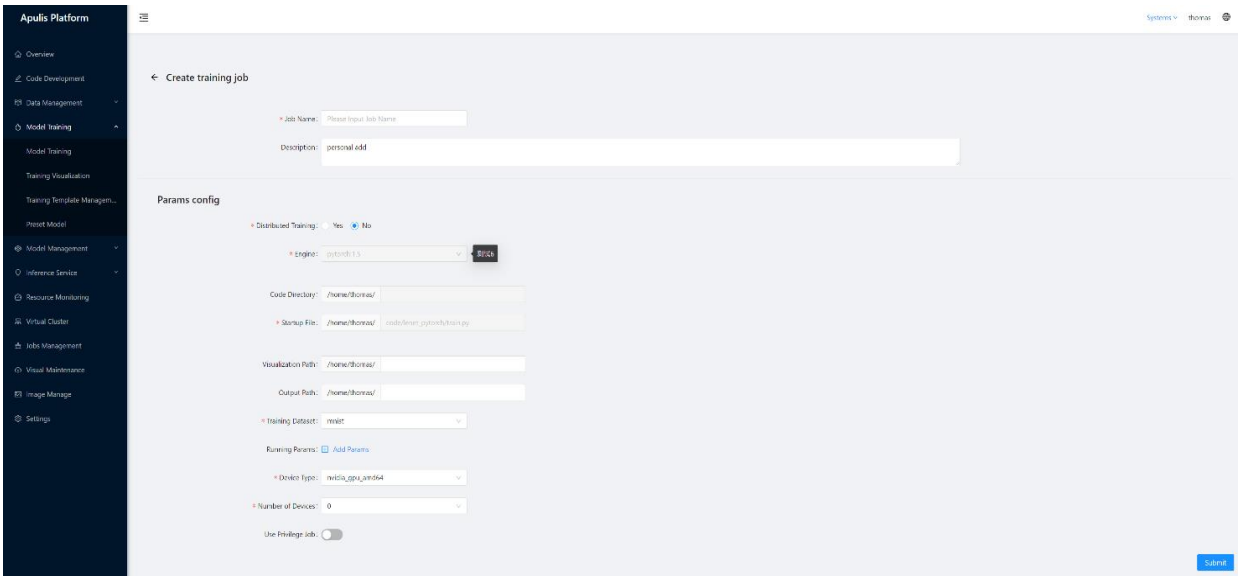


Figure 47-2: Create training job from template

3.6.3 Preset model

The columns of preset model table are as follows:

- Model name: Name of the preset model.
- Model purpose: Description of the model usage and limitation.
- Model accuracy: Accuracy of the preset model.
- Model size: Size of the preset model.
- Creation time: Created time of the preset model.
- Operations: Operations available to the preset model. Currently, it supports “create training job”.
- Search box: you can search the preset model by name (Fuzzy search).

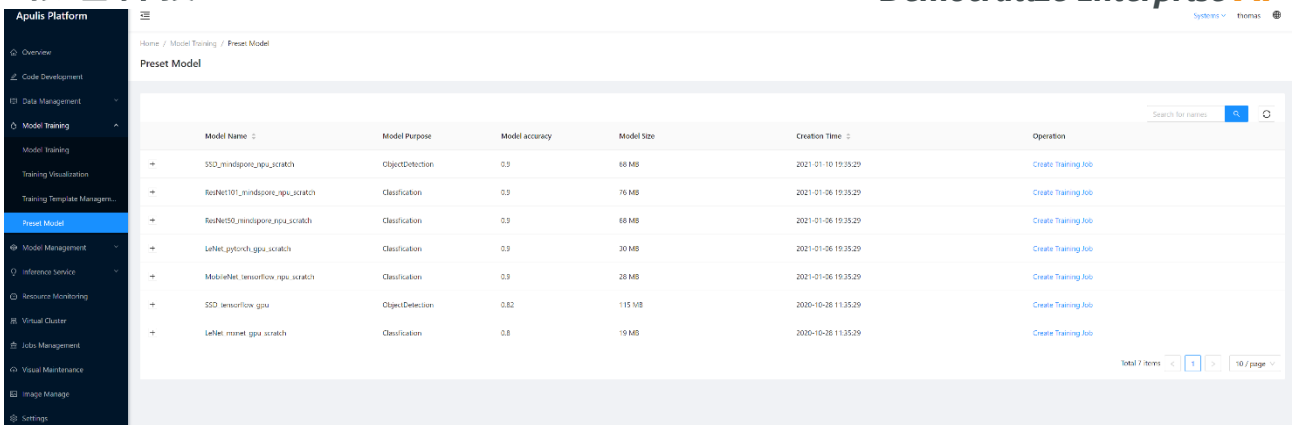


Figure 48: preset model list

The fields you need to create a training job is as follows:

- Job name: required, name of the job. Please note that only alphabetic characters, numbers and underscores are valid input.
- Description: optional, description of the model training job.
- Parameter configuration: required. If you create the job from a template, parameter configuration would be set ready for you and you can modify the parameter as needed.
- Engine: required, algorithm framework.
- Code directory: required, path of the code.
- Startup file: required, it must be a Python file.
- Output path: required, the path where the possible output of the training job should be stored.
- Training dataset: required, dataset for training.
- Run parameters: runtime parameters for your model.
- Distributed training: whether this job is distributed. Distributed job would run on multiple computing nodes, while ordinary job only run on a single computing node.
- Device type: required, type of the requested machine.
- Device Num: required, the number of devices used in training.

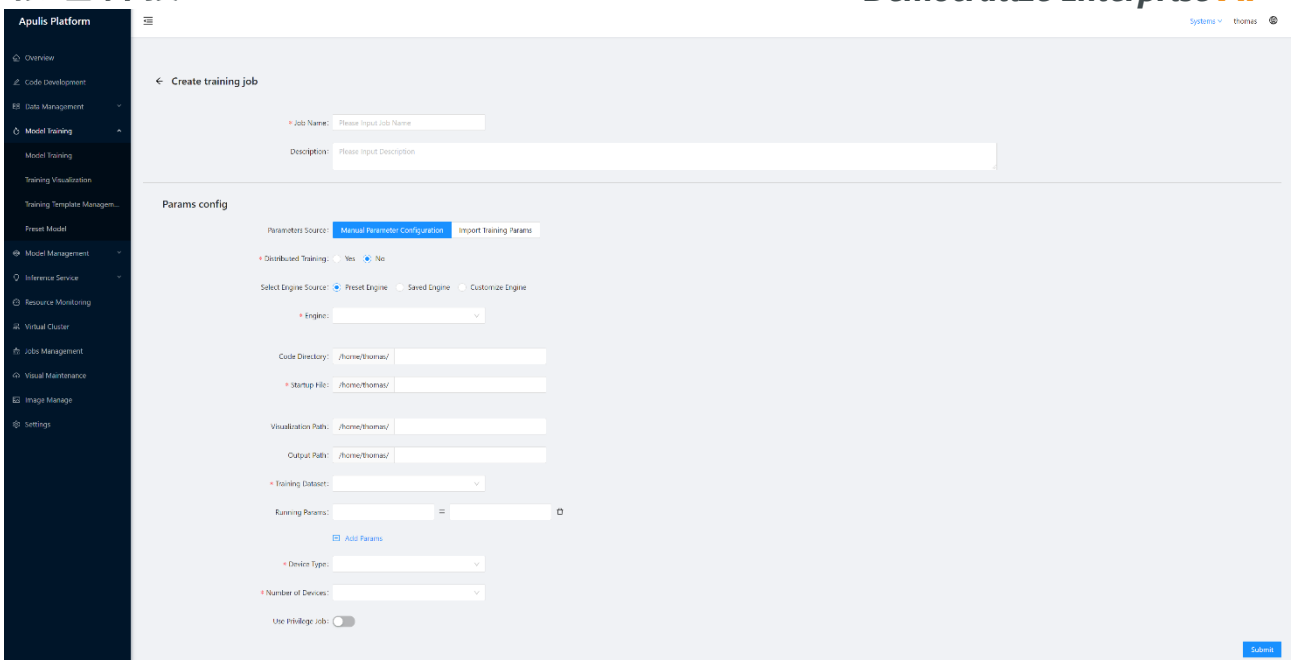


Figure 49: create training job

3.7 Model management

The module of model management consists of 3 sub-menus, namely “My Models”, “Model Evaluation” and “Evaluation Parameter Management”. On “my model”, users can upload, download, delete, and evaluate the model interested. All evaluation jobs created by users are displayed on “Model Evaluation”. The evaluation parameters saved by the user are displayed on “Evaluation Parameter Management”. These saved parameters can be imported whenever the users need to run training tasks using the same parameters.

3.7.1 My models

The page of “My Models” defaults to the “Model Evaluation” page. On the menu of “My Models”, users can create, download, delete, and evaluate. models.

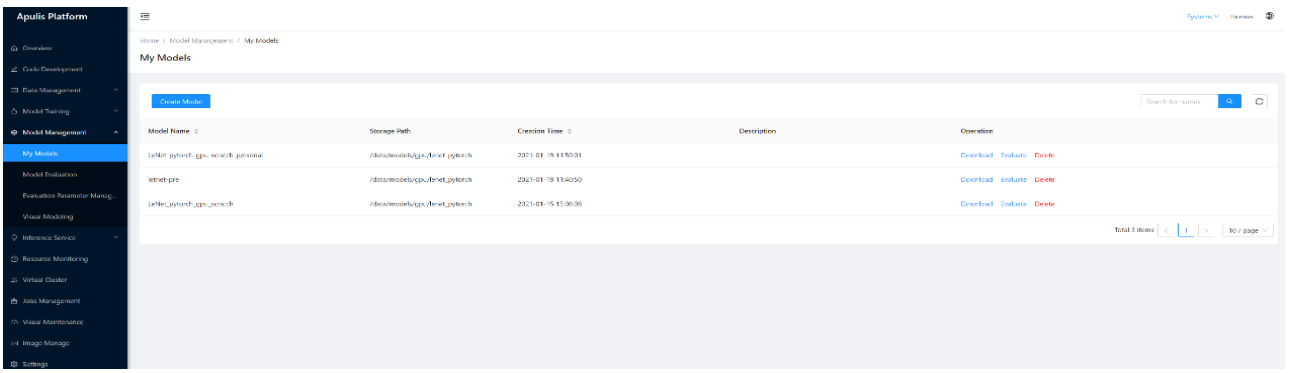


Figure 50: My model

3.7.2 Creating Models

By clicking "Create Models", user enter the page of 'Creating Model'. Model name, description information, model file and model parameter file need to be filled into the form. After entering, "Create Now" would complete the process of creation. There are two ways to set the model file: importing the model file from a completed training job or uploading the model file through the webpage showing below.

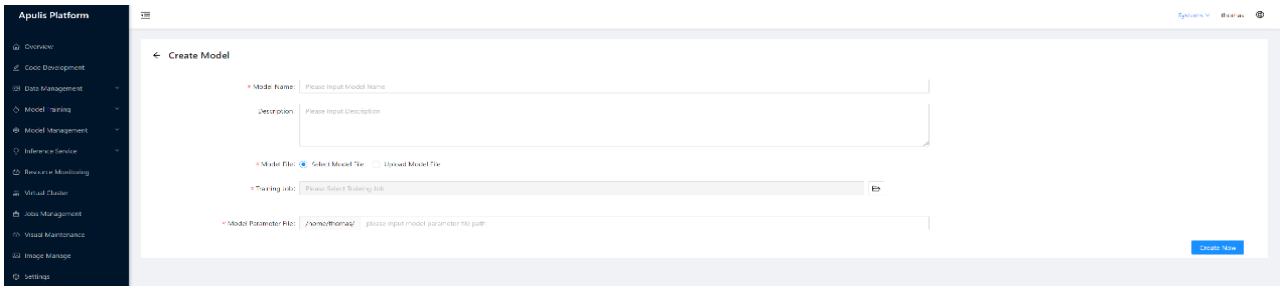


Figure 51: Create model

Model Name: Required, filling in the name of the created model. Please note only alphabetic characters, numbers and underscores are valid.

- Description: Optional, filling in the description of the model.
- Model Parameter File: Required. Entering the path of the model parameter file. The path should exist on the server. If users create a model with a non-existed path, the page will prompt "File or path does not exist".

When the mode of the model file is set to "Select Model File", the training jobs are displayed. By clicking the icon, a pup up window: "Please select training job", would appear and list all completed training jobs, as shown in the figure below. User can select one to continue. After the model is created, the storage path on the list would be refreshed and set to the path of the main code with which user kicks off the training job.

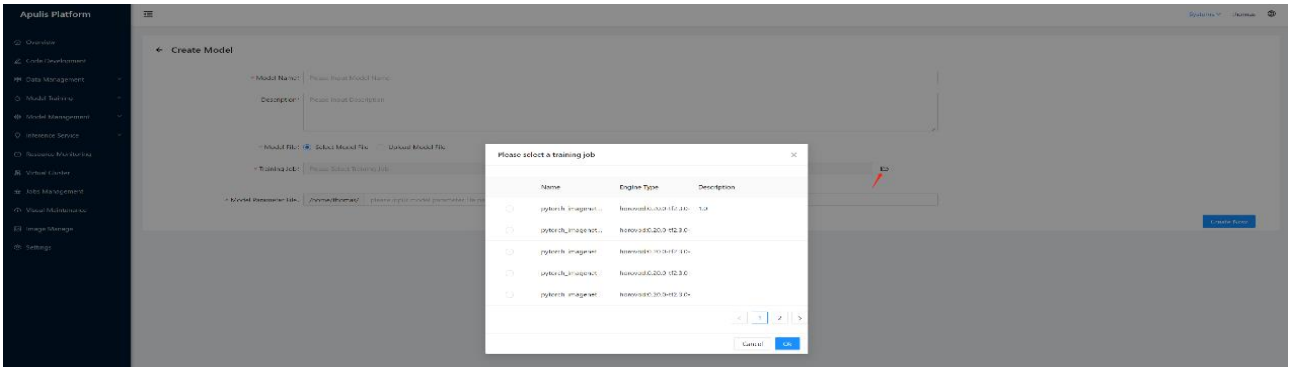


Figure 52: Select training job

The mode of model file is set as "Upload Model File". It would allow users to upload user chosen model file. The supported formats are zip, tar, and tar.gz. After the model is created, the storage path in the model list would be set to the path of the uploaded file.

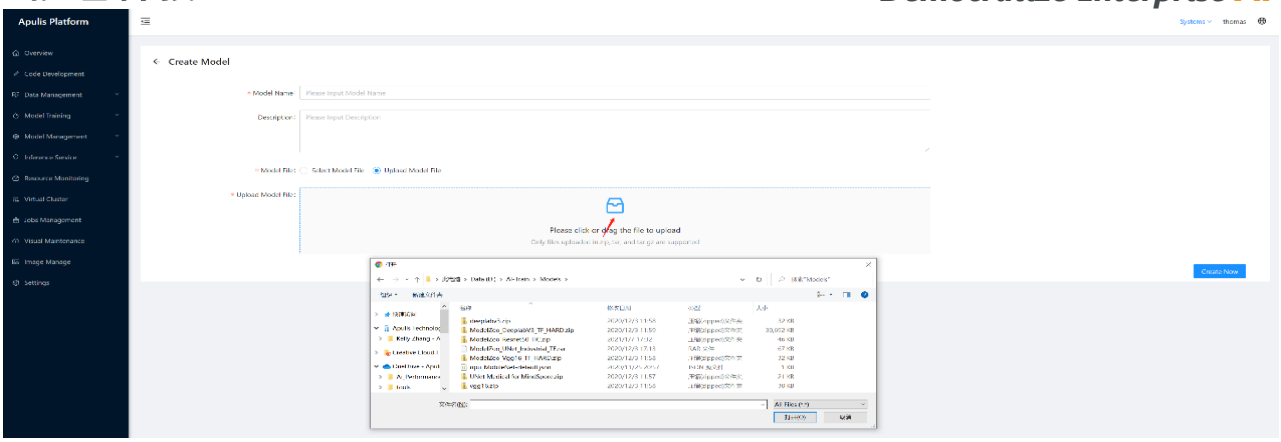


Figure 53: Upload model file

3.7.3 Model list

The model list shows the model name, storage path, created time, description, and operation. More than that, it supports a fuzzy search functionality on model names, as shown in figure 54.

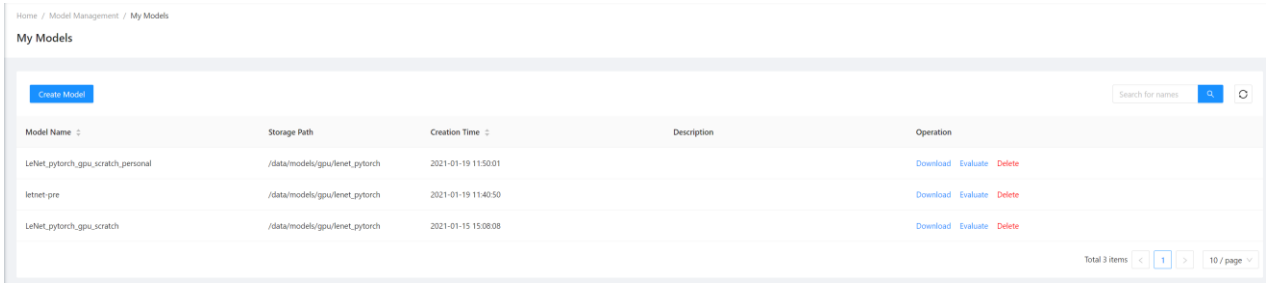


Figure 54: Model List

Model name: Name of the saved model.

Storage path: Path of the saved model. If the model is created by "select model file", this field would show the path of the code user kicks off the corresponding training job. If the model is created by "upload model file", this field shows the path of the uploaded file.

- Creation time: Created time of the model.
- Description: Description of the model or any other information when the model is created.
- Operation: Available operations on the model. It includes model download, deletion, and evaluation. By clicking model download, the file would be compressed and downloaded to the local computer; By clicking delete, the record of model management would be deleted; By clicking model evaluation, it would lead to the page of model evaluation creation.
- Search box: Entering the model name, it would perform a fuzzy search on my model list.

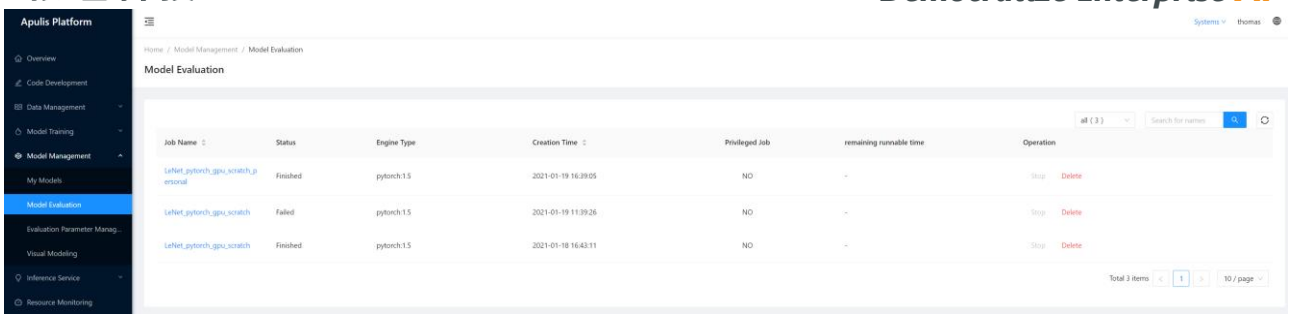


Figure 55: Model Evaluation

3.7.4 Model evaluation

When the model evaluation is created, user cannot change the model name. There are two ways to set up the parameters in the evaluation, first, through the manual configuration of the parameters and second, by importing pre-set parameters. The following items can be configured: engine, code directory, startup file, output path, model parameter file, test data set, running parameters, device type and device number.

Figure 56: Create Model Evaluation

3.7.5 Evaluation list

After the model is created and evaluated, the information below is available.

- Job name: Name of the model evaluation, click to view the evaluation details.
- Status: Status of model evaluation (queuing, scheduling, running, completed, failed, closed).
- Engine type: Engine type selected.
- Creation time: Creation time of the model evaluation.
- Operation: Three states of queuing, scheduling, and running. All three contains a stop buttons, click to stop the evaluation. Stop button is not clickable for completed, failed, and closed, since they are closed status.
- Status filtering: Select a status to display corresponding list.
- Search: Enter the job name to do fuzzy search on the evaluation list.

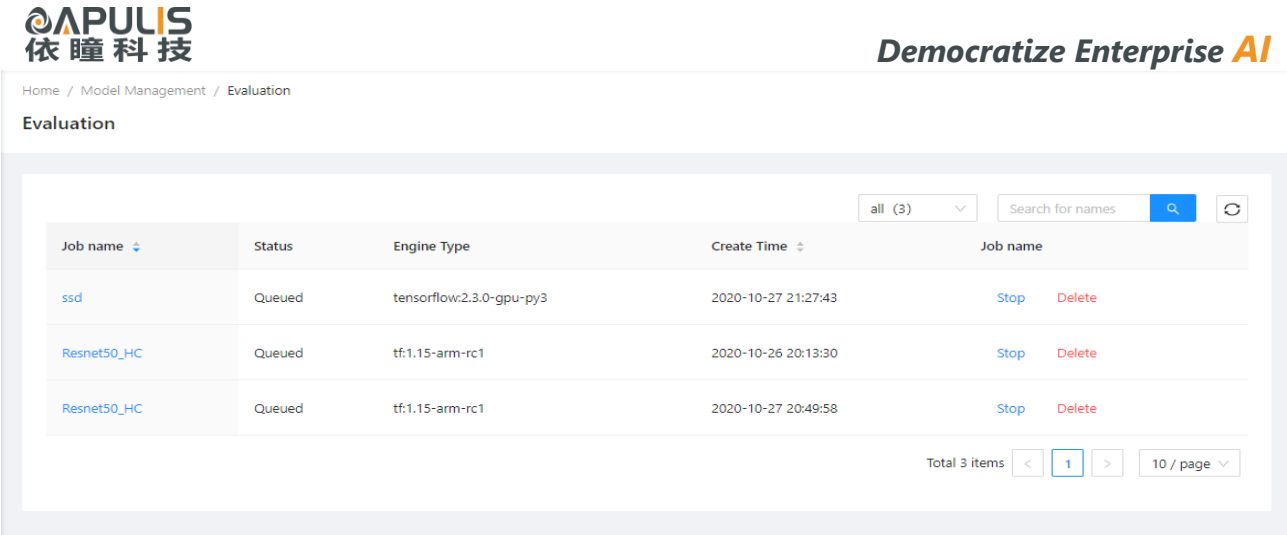


Figure 57: Evaluation List

Evaluation Detail:

Details: Detailed information about the model, including model name, created time, evaluation status, engine type, test data set, code directory, startup file, output path, model parameter file, device type, device number, and running parameters.

Obtain evaluation results (evaluation breaks into two categories: detection and classification. the following figure is detection):

- Accuracy: Accuracy of model evaluation.
- Precision: Precision of model evaluation.
- Recall: Model recall rate.
- Recall_5: Top 5 model evaluation recall rate.

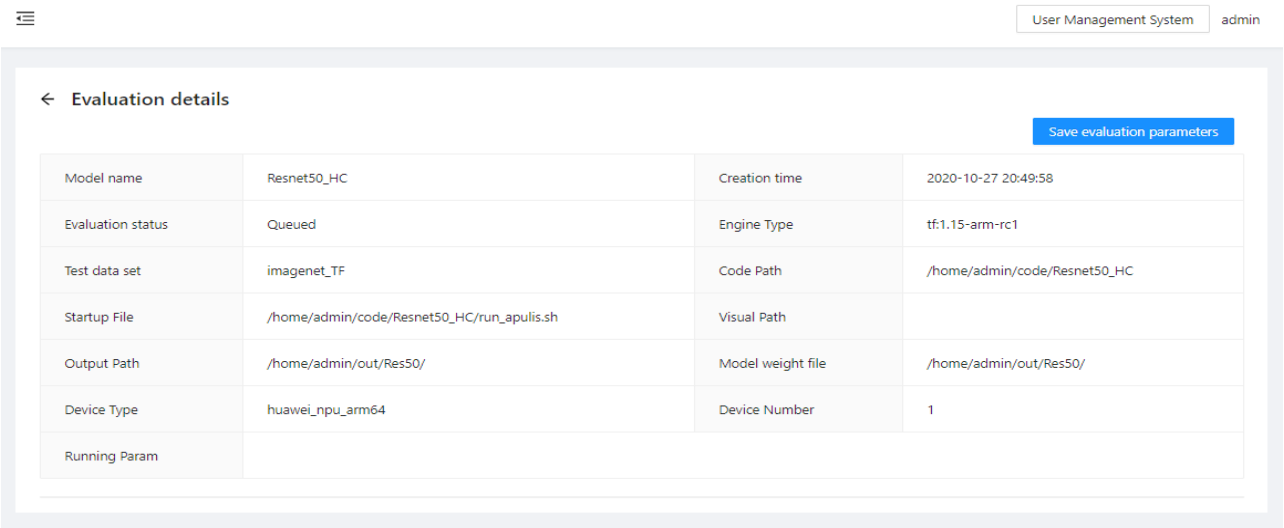


Figure 58: Evaluation detail(detection)

Obtain evaluation results (evaluation breaks into detection and classification, the following figure is classification):

- Classification_Loss: Classification loss.
- Localization_Loss: Regression loss.
- Regularization_Loss: Regularization loss.
- Total_Loss: Total loss.
- mAP: Average.

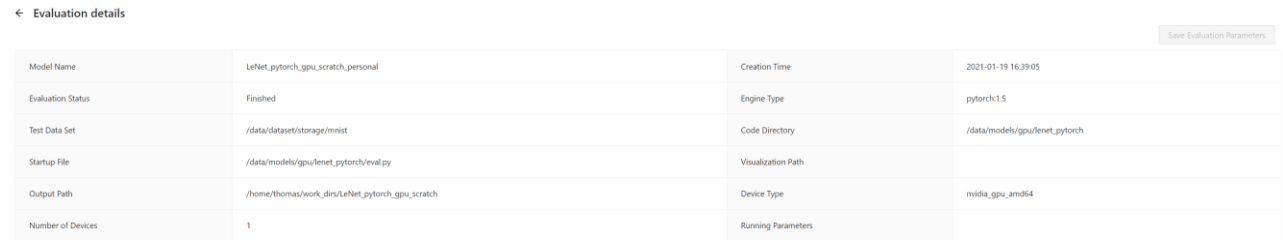


Figure 59: Evaluation detail(classification)

Save evaluation parameters (after saving, users can save the model parameters to the evaluation parameter management, and users can directly import those parameters when creating the model evaluation):

- Configuration name: Required, name of evaluation parameter. Please note only alphabetic characters, numbers and underscores are valid.
- Type: Required, it cannot be modified.
- Engine type: Required, it cannot be modified.
- Description: Required, description of the evaluation parameter.

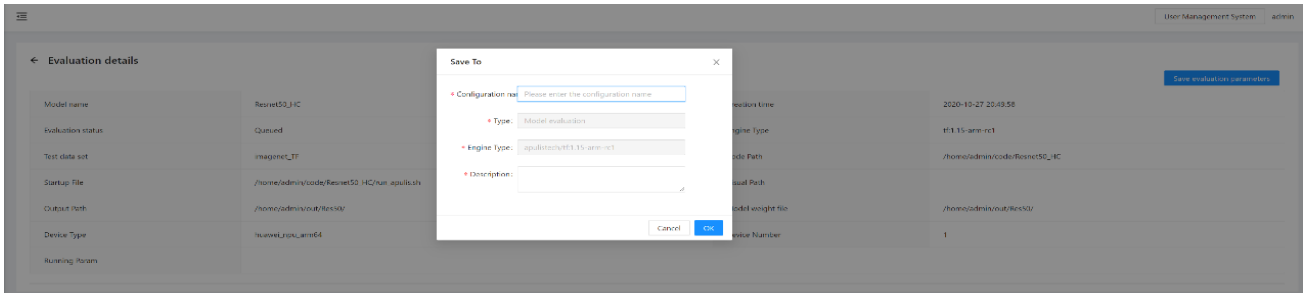


Figure 60: Save Evaluation Parameters

3.7.6 Evaluation parameter management

Evaluation parameter name: Name of the evaluation parameter.

- Engine type: Type of engine used to evaluate the parameter.
- Creation time: Creation time of the evaluation parameter.
- Description: Description of the evaluation parameter.
- Operation: Edit (edit parameter), Delete (delete parameter).
- Search: Enter evaluation parameter name for fuzzy search.

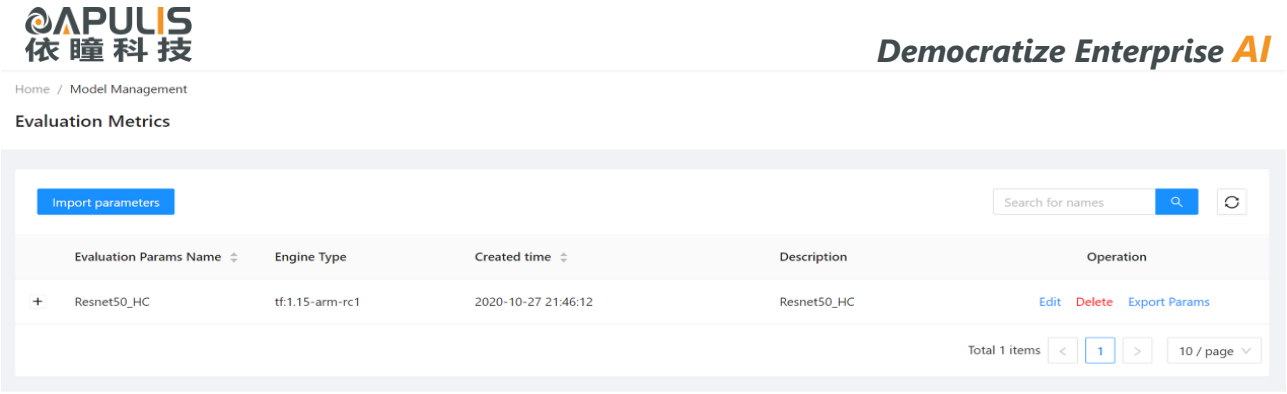


Figure 61: Evaluation parameter list

Edit evaluation parameters:

The evaluation parameters can be edited. Editable parameter includes: Evaluation parameter name, Description, Engine, code directory, startup file, output path, Model parameter path, training dataset, operating parameters, device type, and device number.

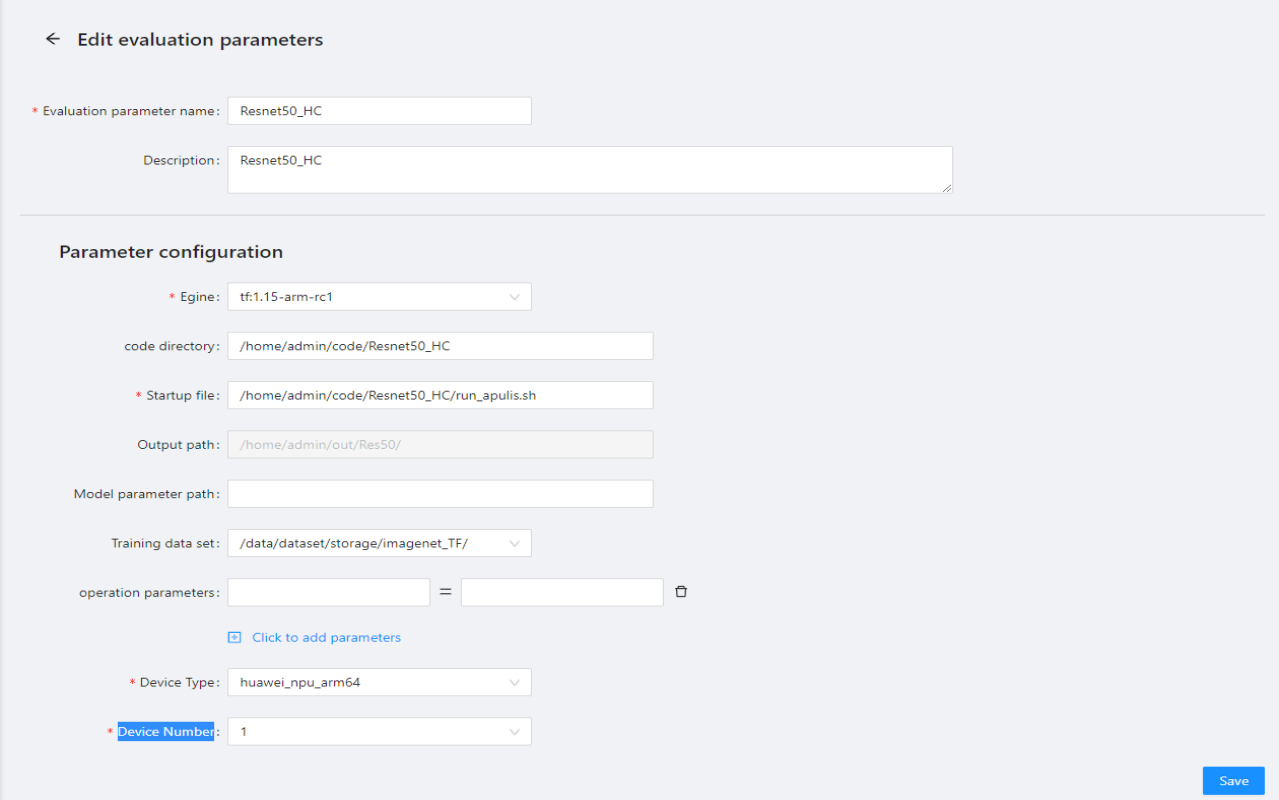


Figure 62: Edit evaluation parameters

3.8 Inference Service

Users can use a trained algorithm model to ~~Chuangjing~~create inference operations. The platform supports cloud inference mode and edge inference mode.

3.8.1 Create Cloud Inference Job

Click "Inference Service" – "Central Inference" – "Create Inference Job" in the menu bar. This leads to a page to create a Cloud Inference Job, as shown in Figure 63.

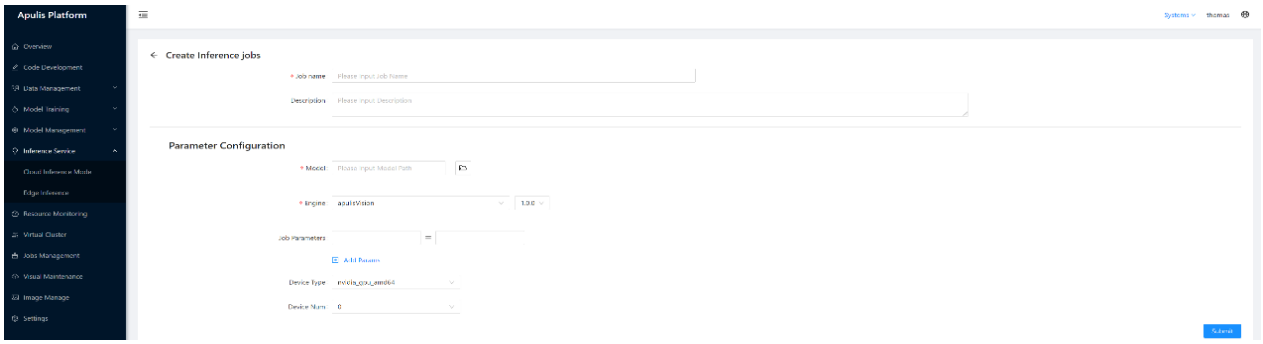


Figure 63: Inference Job Creation Page

The form to create a cloud inference job consists of seven items: Job name, Description, Model, Engine, Job Parameters, Device Type, and Device Num.

- Job name: Required, name that identifies the inference job.
- Description: Optional, description of the inference job.
- Engine: Required, current choices are: 1) Tensorflow-1.15.0 and 2) Mindspore-1.1.0.
- Model: Required, enter the directory address of the algorithm model that has been trained.
- Job Parameters: Optional, inference job parameters that need to be specified.
- Device Type: Required, current supported choices are: 1) CPU and 2) GPU.
- Device Num: Required, default value is 0.

3.8.2 Inference job management

This window displays a list of inference jobs that are created by the current user, as shown in figure 64.

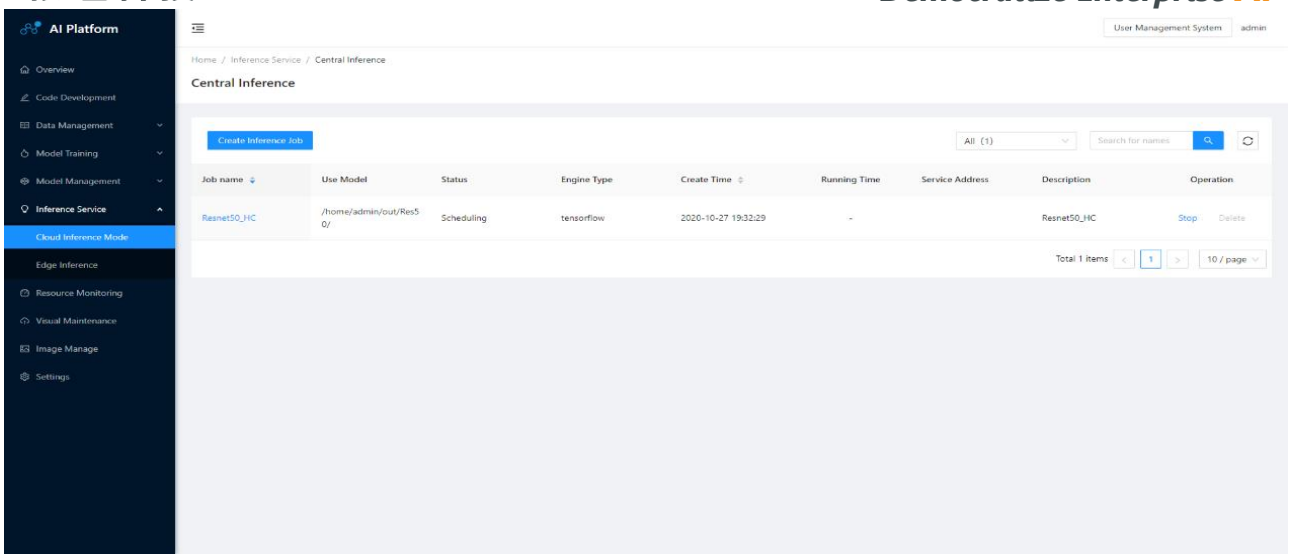


Figure 64: Inference job management

The window shows a list of inference jobs, each of which includes Job Name, Model, Status, Engine Type, Creation Time, Running Time, Service Address, Description, and Operation.

- Job Name: Name defined by the user when the inference job is created.
- Model: Model file URL or address.
- Status: Status of the current inference job.
- Creation Time: Creation time of inference job.
- Running Time: Duration of the inference job, if the inference job is closed, the duration will not increase.
- Service Address: URL of the inference job.
- Description: Description of the inference job.
- Operation: Inference job can be stopped and/or deleted. Currently, we may not restart a stopped inference job. To restart, users need to create a new inference job.

3.8.3 Inference Job: WebUI

Click on the job name to open inference job in the running state and enter the page to manipulate inference jobs, as shown in figure 65.

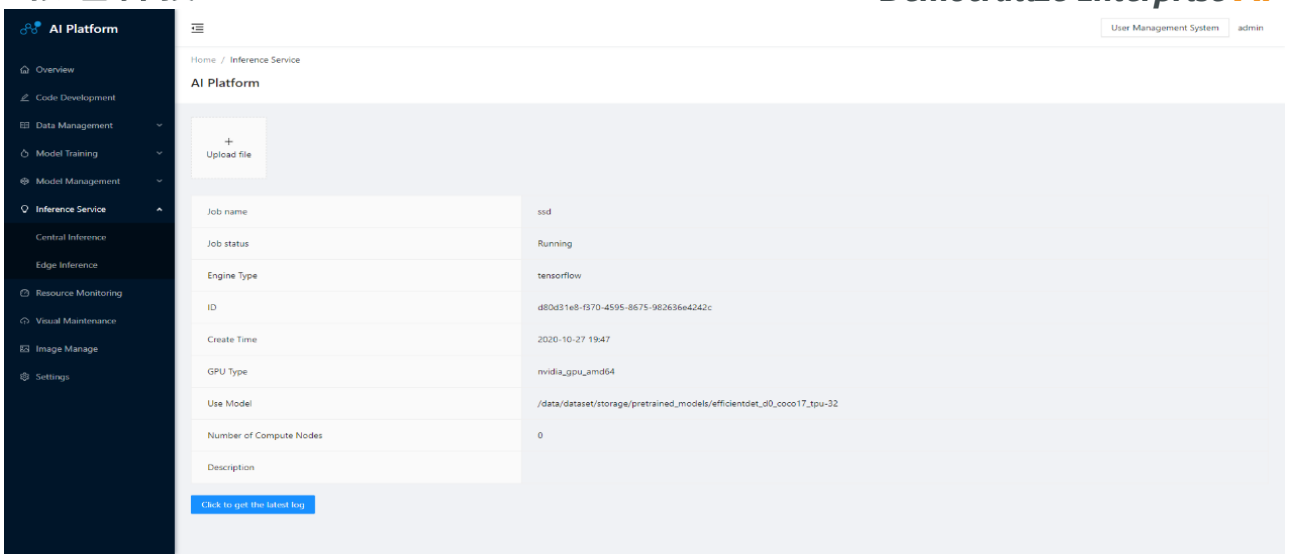


Figure 65: Use Inference Assignments

Click to upload a picture, as shown in figure 66.

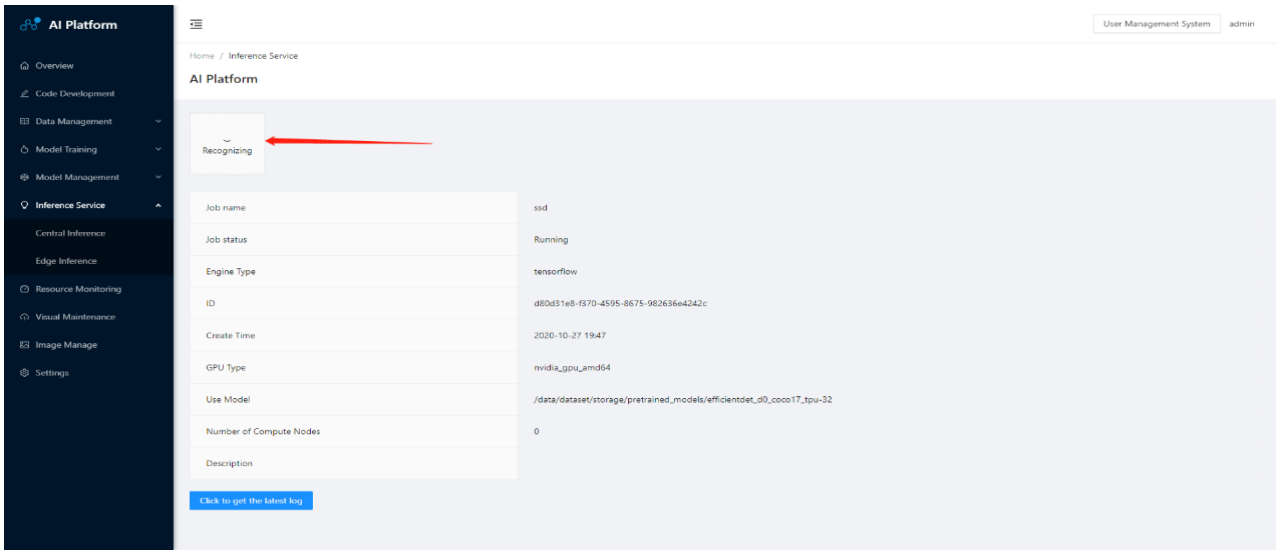


Figure 66: Upload Pictures

Click Start Recognition and obtain the result.

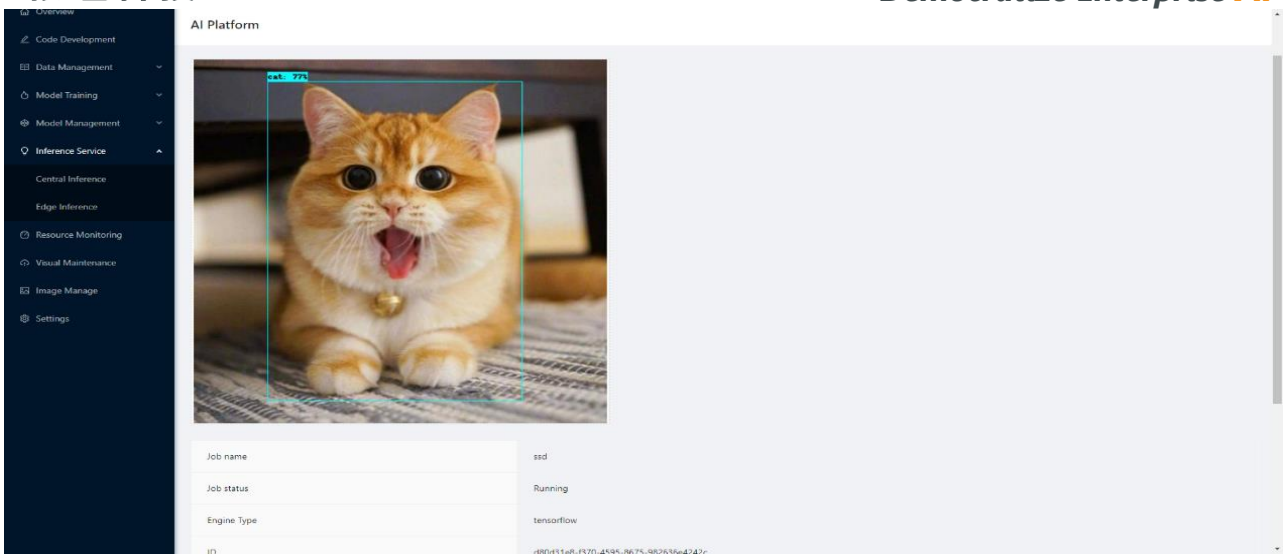


Figure 67: Recognition Result

3.8.4 1.2.4 Create Edge inference job

Click [Inference Service] - [Edge Inference] - [New Inference] in the menu bar, to create a new edge inference, as shown in figure 68.

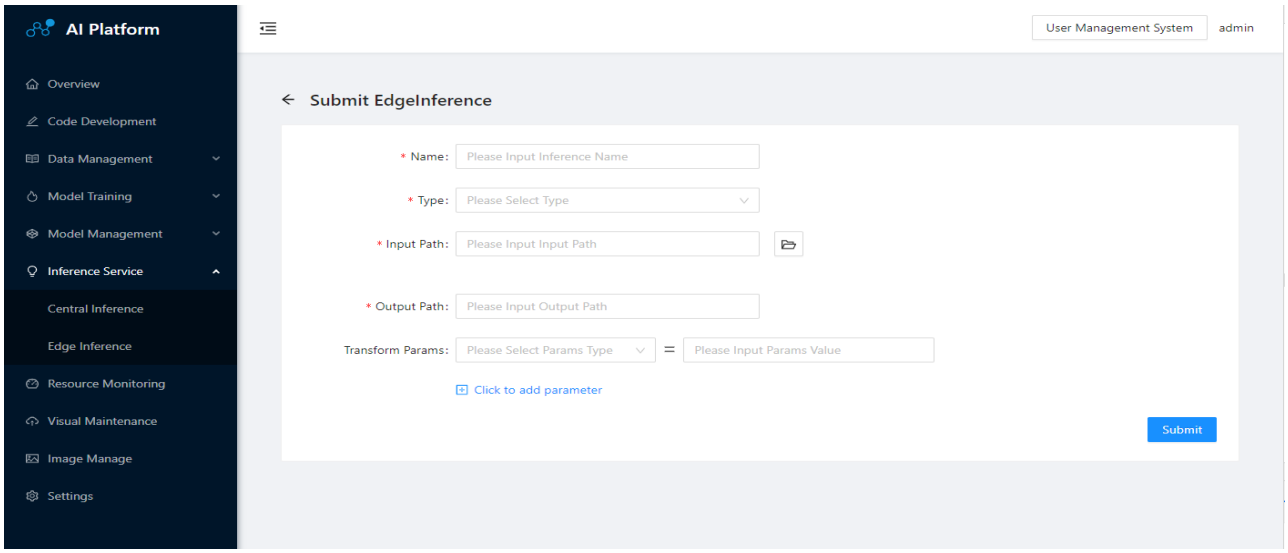


Figure 68: New Inference Pop-up Window

To create a new edge inference job, four items are needed including Inference Job Name, Type, Input Path, and Output Path.

- Inference Job Name: Required, users can customize input.
- Type: Required, a convertible type supported by the system.
- Input path: Optional, Input path where the model file is located.
- Output path: Optional, Output path of the converted model file.

3.8.5 1.2.5 Setting up FD Server

Click “Settings”; A window to set the FD (Huawei Fusion Director) server will pop up.

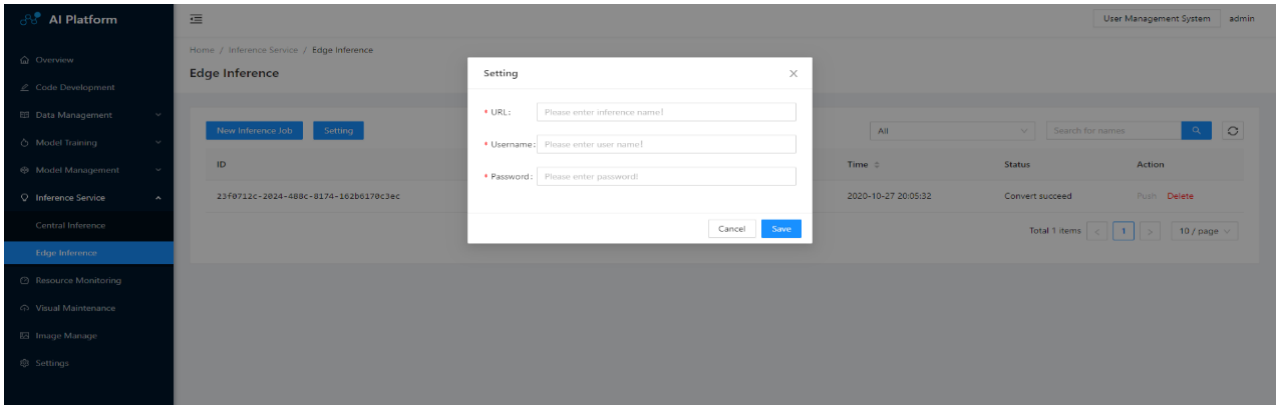


Figure 69: Setting up Inference Server

3.8.6 1.2.6 Push Model to FD Server

Users can select a model (properly converted for FD server deployment) and push it to the FD server, as shown in figure 70.

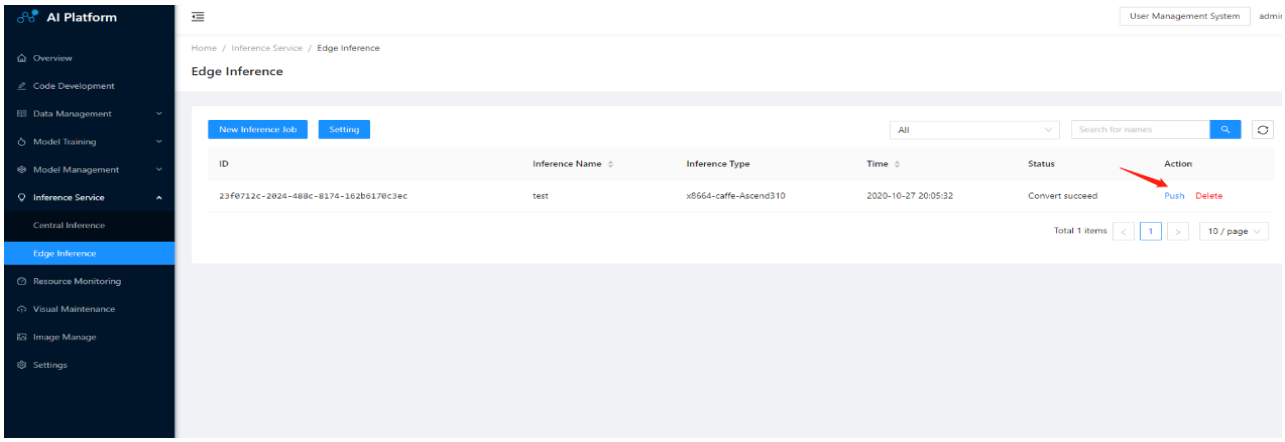


Figure 70: Pushing Model

3.9 1.3 Resource monitoring

3.9.1 Check VC usage

Users can check each VC resource usage through [VC Device usage] on the top left corner.

Figure 71: Check VC Device usage

3.9.2 Check Cluster Usage

Users could view CPU or NPU usage status of the worker node in the cluster by selecting [device usage] on the top left corner.

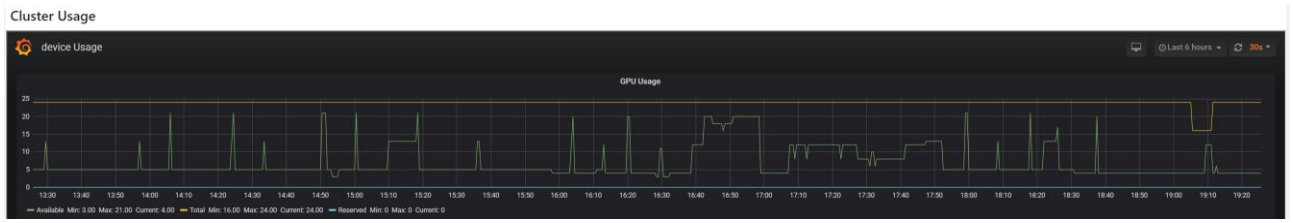


Figure 72: View cluster device usage

3.10 Virtual cluster

[Virtual Cluster] could only be accessed by operation and maintenance administrator. Administrators can create, delete, edit virtual clusters, allocate virtual cluster resources, and manage virtual clusters for users.

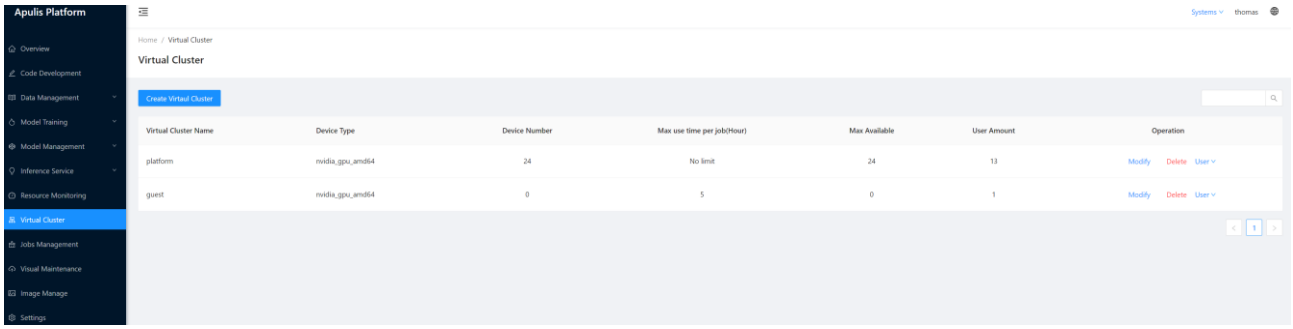


Figure 73: List of Virtual Clusters

3.10.1 Create a Virtual Cluster

To set the name of the virtual cluster, the number of devices, the maximum number of users and the maximum use time (unit: hour), users could
click [Create Virtual Cluster] on the upper left corner and then click [OK] on the lower right corner.

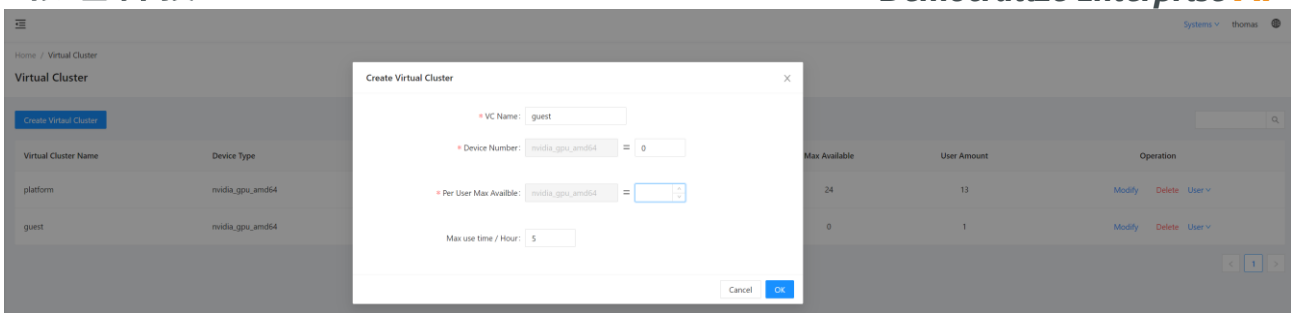


Figure 74: Create a Virtual Cluster

3.10.2 Delete Virtual Cluster

Select [Delete] on the right side of the virtual cluster and click [OK] in the pop-up dialog window to delete virtual cluster.

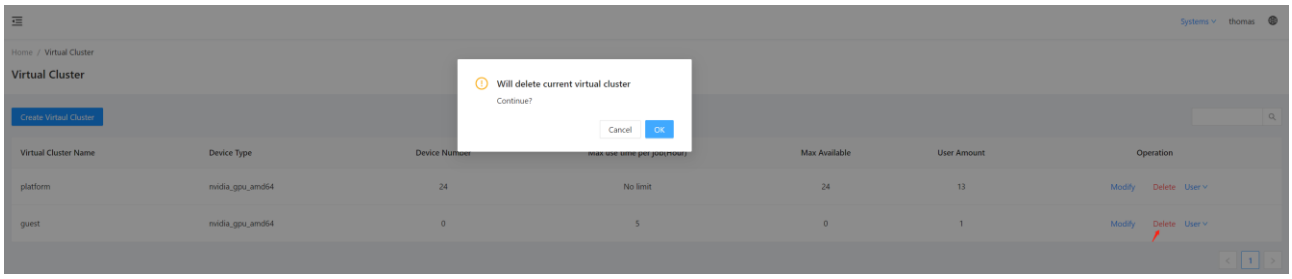


Figure 75: Delete Virtual Cluster

3.10.3 Relate User

Click [Relate User] on the right side of a virtual cluster. In the pop-up dialog window,

administrator can choose to add a user created in the user management platform to the virtual cluster, and then the user has resources of the virtual cluster they assigned to.

Figure 76-1: Relate a user

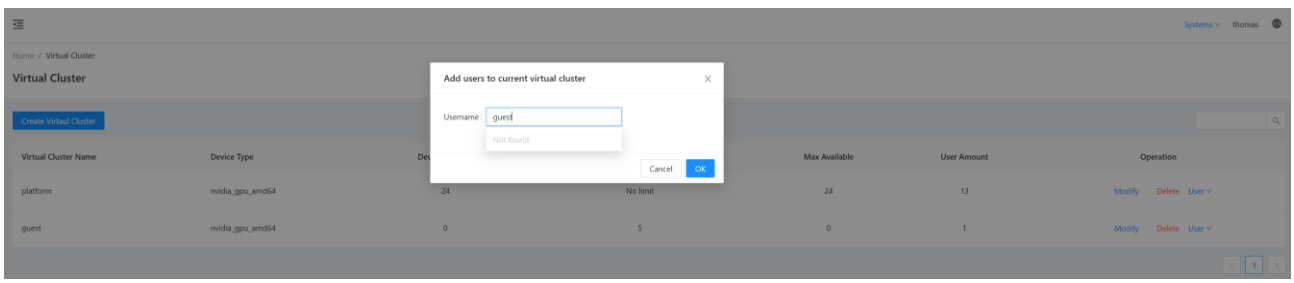


Figure 76-2: Relate a created user

3.10.4 View users

Click [View User] on the right side of a virtual cluster, and operator can view all users added to the virtual cluster in the dialog window.

Figure 77: View the associated cluster users

3.11 Task management

Administrator could use the admin user account to login to the Apulis artificial intelligence platform. By clicking "Job Management" in the left menu bar, the administrator can see the jobs run by all users on the current platform, and they can "stop" the tasks in the task list. After a task stops, the occupied NPU resources will be released.

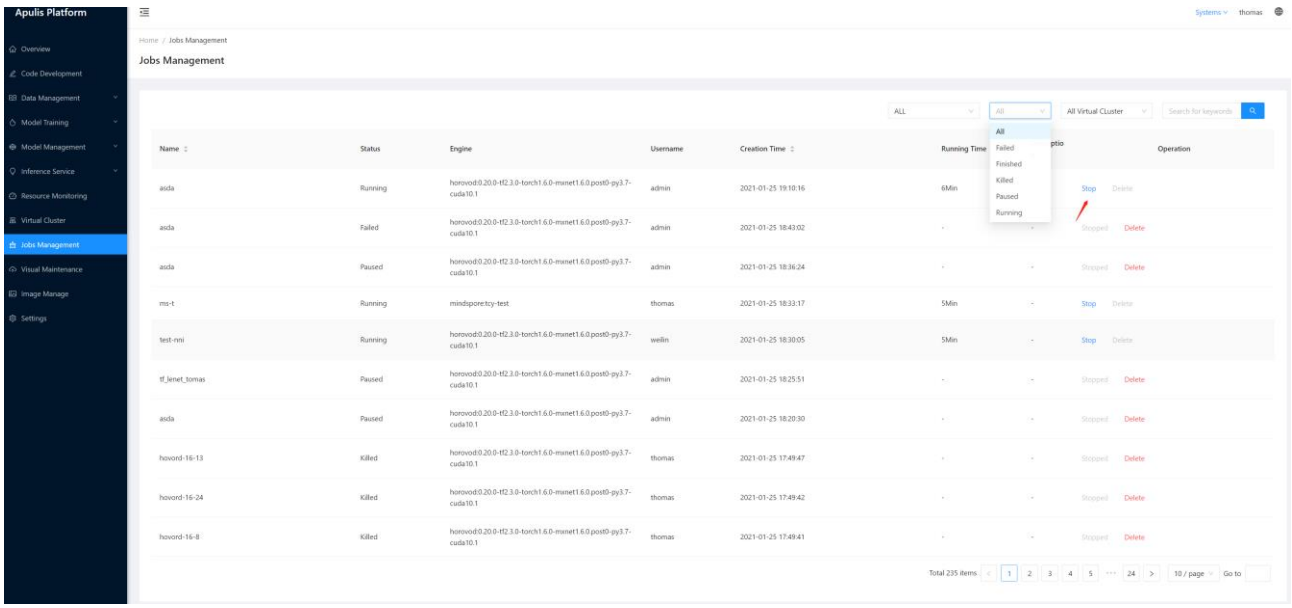


Figure 78-1: Admin User Stops Jobs

Filtering tasks: Users can filter the task list by type, status, virtual cluster and keywords; The figure below shows an example by searching key word “tf” in the search bar;

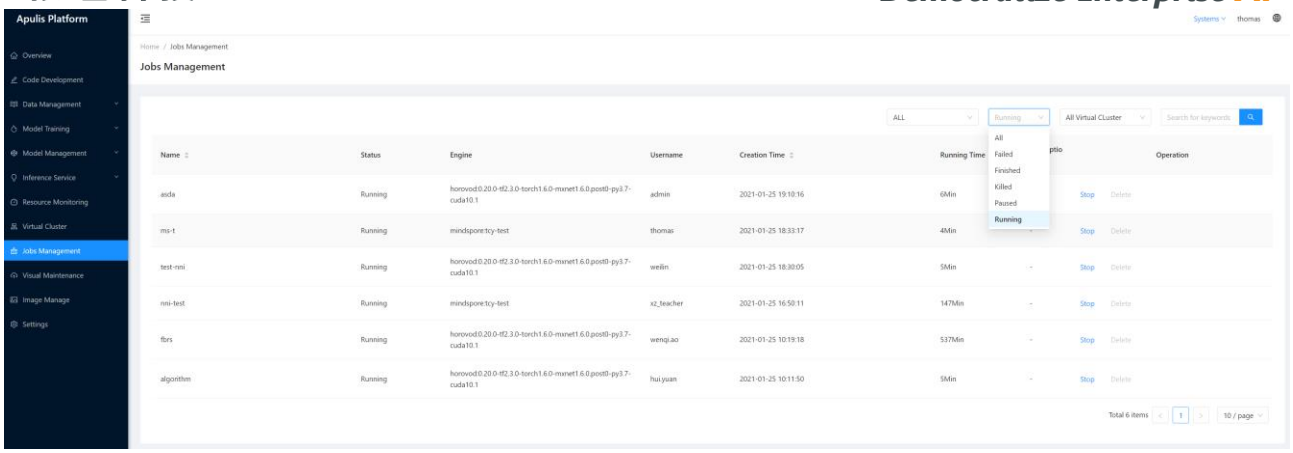


Figure 78-2: Filter tasks

Note: Non-admin users do not have access of this functionality.

3.12 Visual Maintenance

3.12.1 View Alert log status

The platform is configured with storage usage alert by default, and the monitoring status can be viewed in the [Alert rules] window.

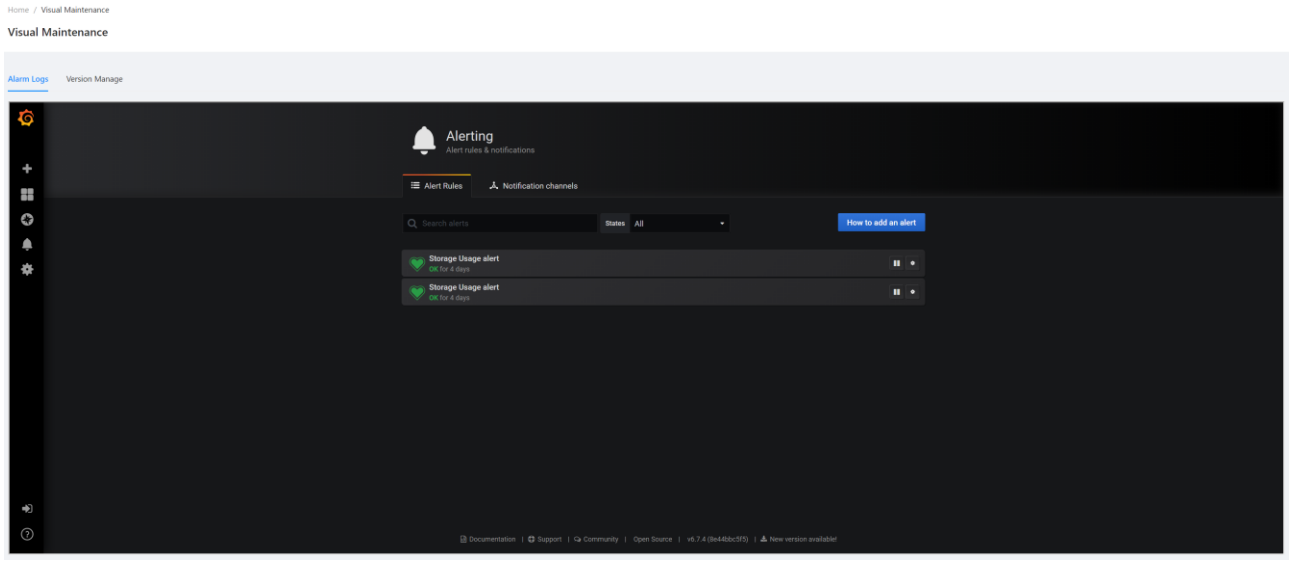


Figure 79: View Alert log status

3.12.2 Configure Notification list

Users can configure the email list for those who needs to receive alert information, and the account in the email list will receive the corresponding alert information.

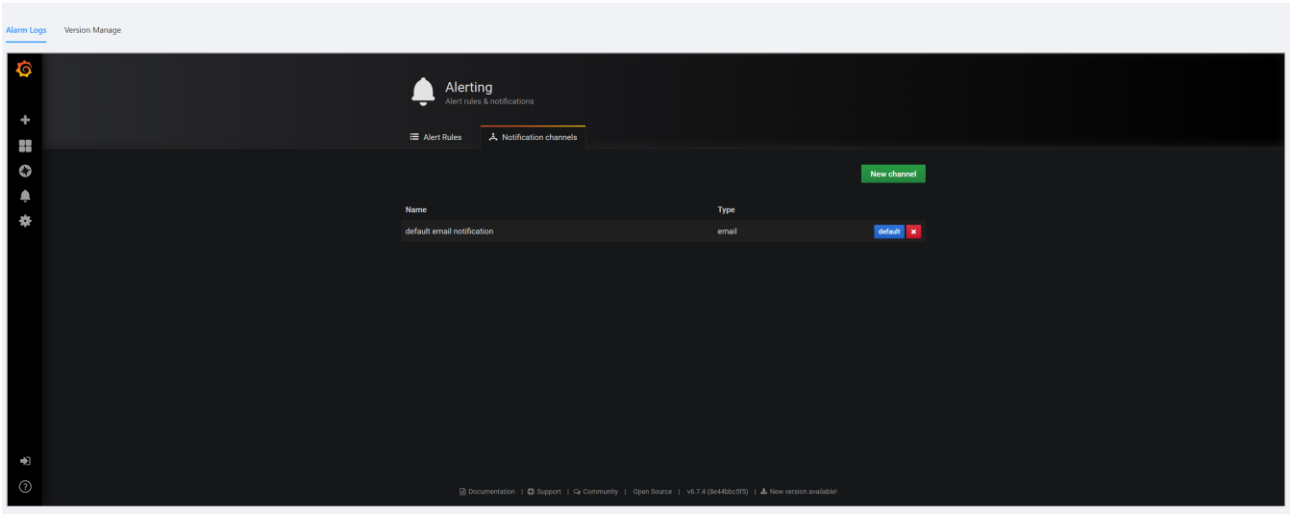


Figure 80: Edit or delete Notification list

3.12.3 Version management

Click [Version Management] on the right side of the window to view information of current version.

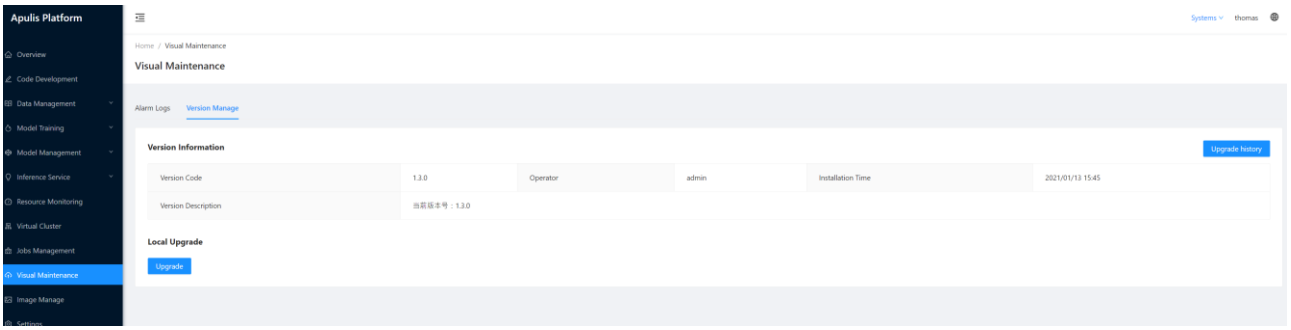


Figure 81: Version Management

3.13 Image management

After the user saved the image in the code development environment, users can view or delete the saved image in the image management window. For the method of saving the image, please refer to section 3.4.7.

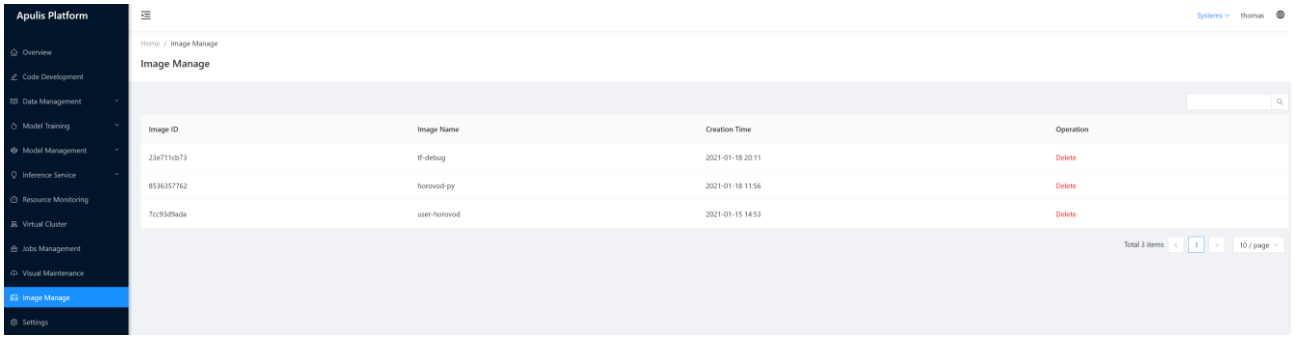


Figure 82: Image management

3.14 Settings

In the settings window, users can set the time interval for platform data refreshment, so that they could obtain the task status in time. When a user occupies associated with current cluster resources, they could view the virtual cluster currently in used; When the user is associated with multiple virtual clusters, they could switch virtual cluster in Settings page to set up different resource configurations.

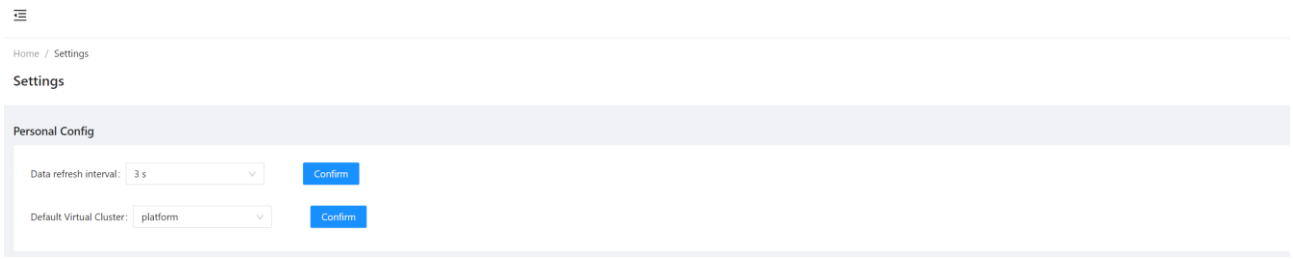


Figure 83: Image management

4 User Management System

The user management system page includes the top menu bar, the console, and the admin page, which includes users, user groups, and roles.

4.1 Dashboard

Displays the number of existing users, user groups, and roles. Click [Create user] to go to the user creation page; click [Create user group] to go to the user group creation page; click New role to go to the role creation page.

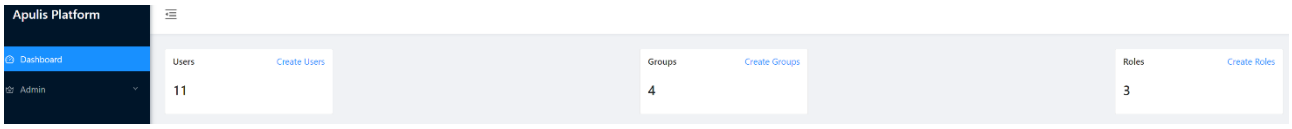


Figure 84: Dashboard

4.2 Admin Page

The page includes users, user groups, and roles. User includes user list and new user, group includes group list and new group, role includes role list and new role. Only administrators can access admin page. Administrators can associate roles to users to restrict their access permissions on the platform.

4.2.1 User

User page includes [user list] and [new user]. User list displays the existing users. Click “New user” to create a new user.

4.2.2 User List

The user list includes username, nickname, phone number, email, and description.

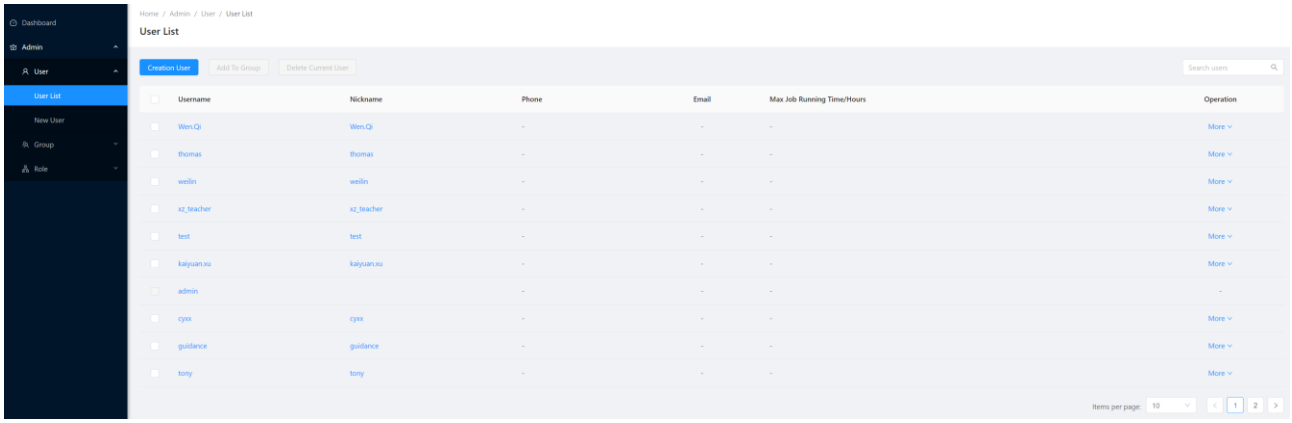


Figure 85: User list

- Username: Username, this field is unique and cannot be repeated.
- Nickname: User's nickname.
- Phone: Phone number.
- Email: Email information.

- Operation: Operation includes modifying user's role, associating the user to specific user group and deleting the user.
- The default admin account: Admin, the account only allows password modification, but it cannot delete or modify user's roles. Newly created administrator account could do all user manipulation.

Click the username or nickname in the list to go to the user details page.

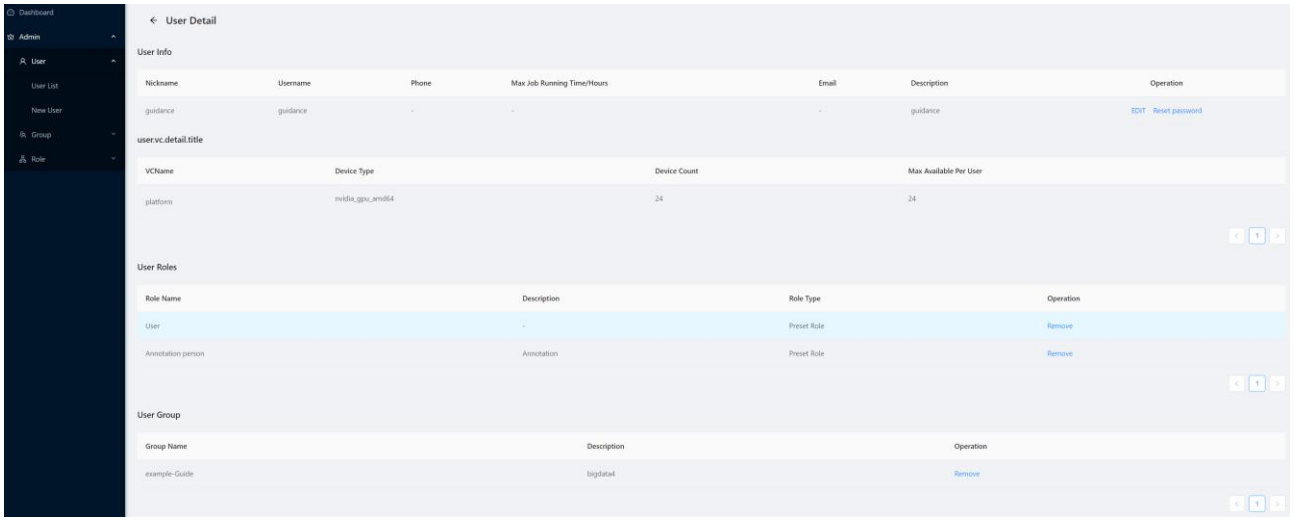


Figure 86: User details page

4.2.3 New User

To create a new user please enter user information -> associated role -> confirm.

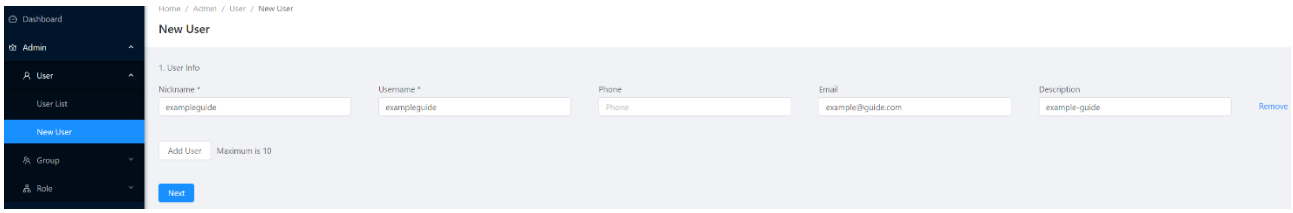


Figure 87: Create User: Enter user information

Nickname and username are required fields. After entering, click next to go to role association page. Administrators could add at most 10 users at a time.

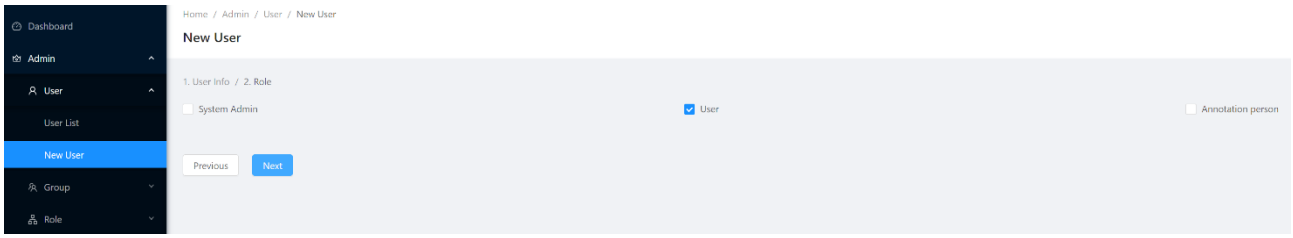


Figure 88: Create User: Associating roles

When associating roles, users need to select at least one role association and click [Next] to go to the preview page or click [Previous] to return to the previous step.

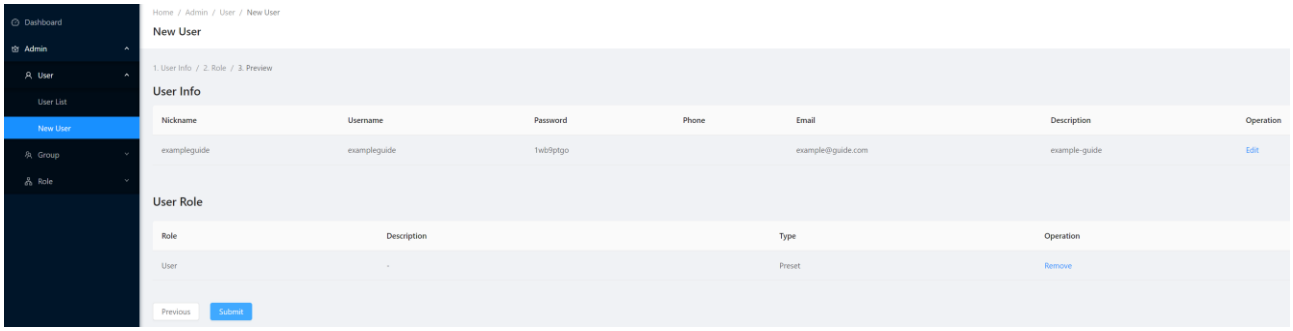


Figure 89: Create User & Confirm

The page displays information of the new user; Click [Submit] to create a new user; Click [Previous] to return to the previous step.

4.2.4 Group

Group includes [group list] and [new group]. The group list displays the list of existing user groups. Click New Group to create a new user group.

4.2.5 User Group List

The user group list includes group name, description, creation time, and operation.

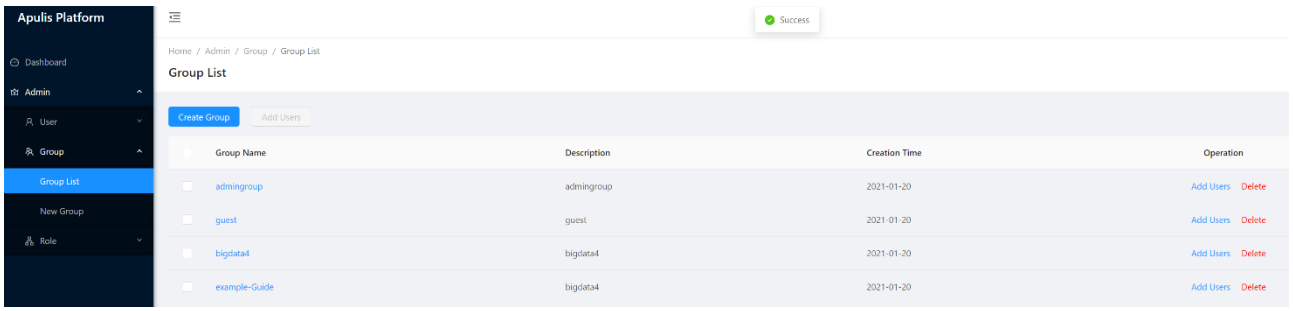


Figure 90: Group list

- Group name: This field is unique and cannot be repeated.
- Description: Description of the user group.
- Creation time: Creation time of the user group.
- Operation: Add users to or delete users from user group

Click the group name in the list to jump to the group details page, which displays the group information, the roles that user group contains, and users in the user group.

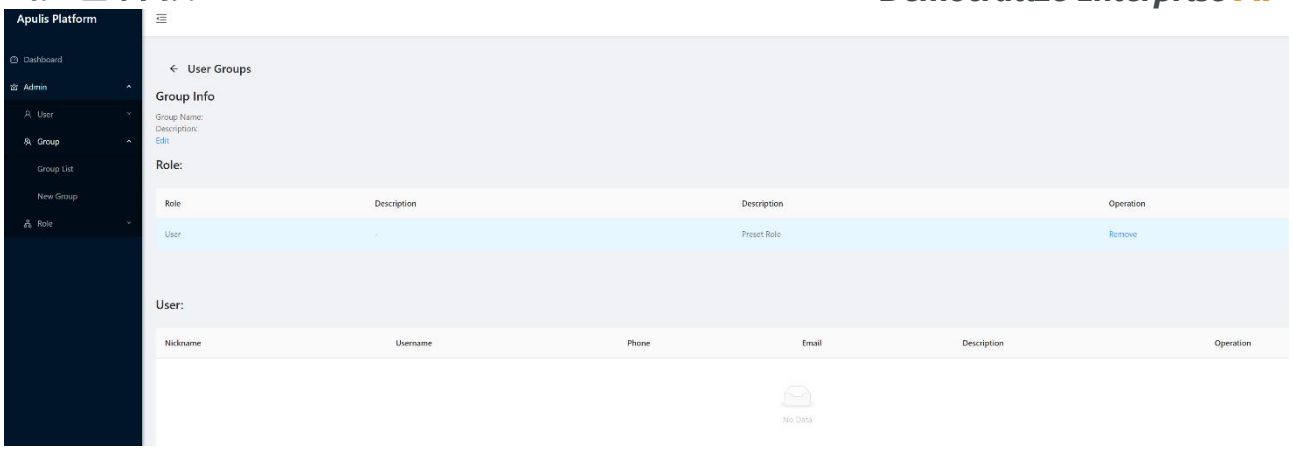


Figure 91: User group details page

4.2.6 New Group

It takes 3 steps to create a group, enter group information -> relate role -> confirm.

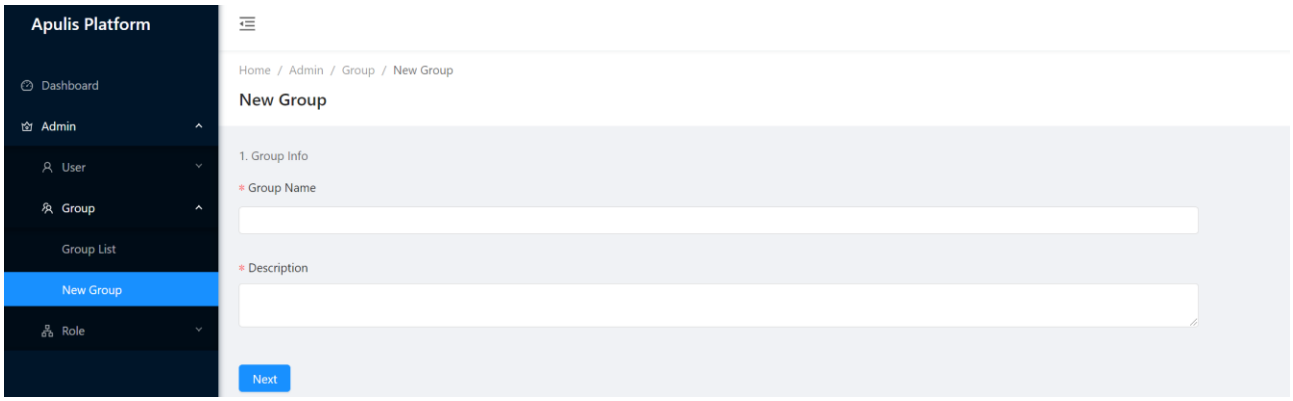


Figure 92: Create Group: Group Information

Group name and description are required fields. After entering, click Next to go to page of role association.

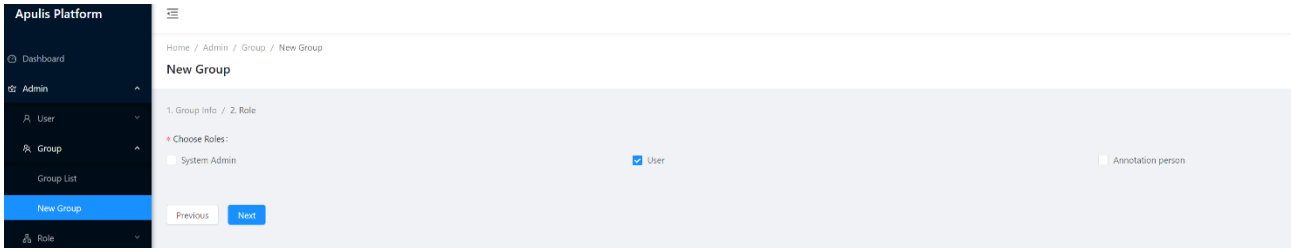


Figure 93: Create Group: Associating Roles

At least one role association must be selected, click Next to jump to the confirmation page, or click Previous to return to the previous step.

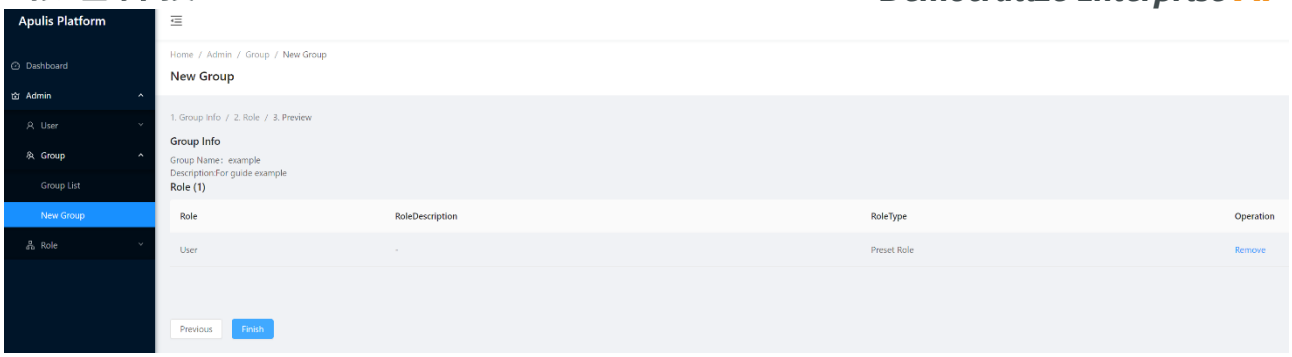


Figure 94: Create Group: Confirm

Shows the new user group related information, click submit to submit creation of a new user group, click previse to return to previous step.

4.2.7 Roles

It includes [role list] and [new role], the role list displays the existing roles, click [new role] to create a new role.

4.2.8 Role List

The role list includes role name, description, type, and operation.

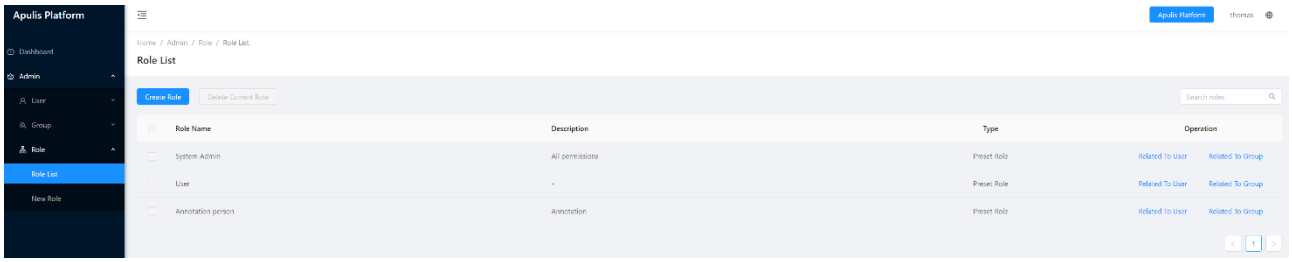


Figure 95: Role list

- Role name: This field is unique and cannot be repeated.
- Description: Description of the role.
- Type: Role types could be break into preset roles and user-defined roles. The preset role is created by the system by default and cannot be deleted.
- Operation: Relate users, relate user groups, and delete roles.

4.2.9 Create Role

It takes 3 steps to create a role, enter the role name -> description -> level of permission.

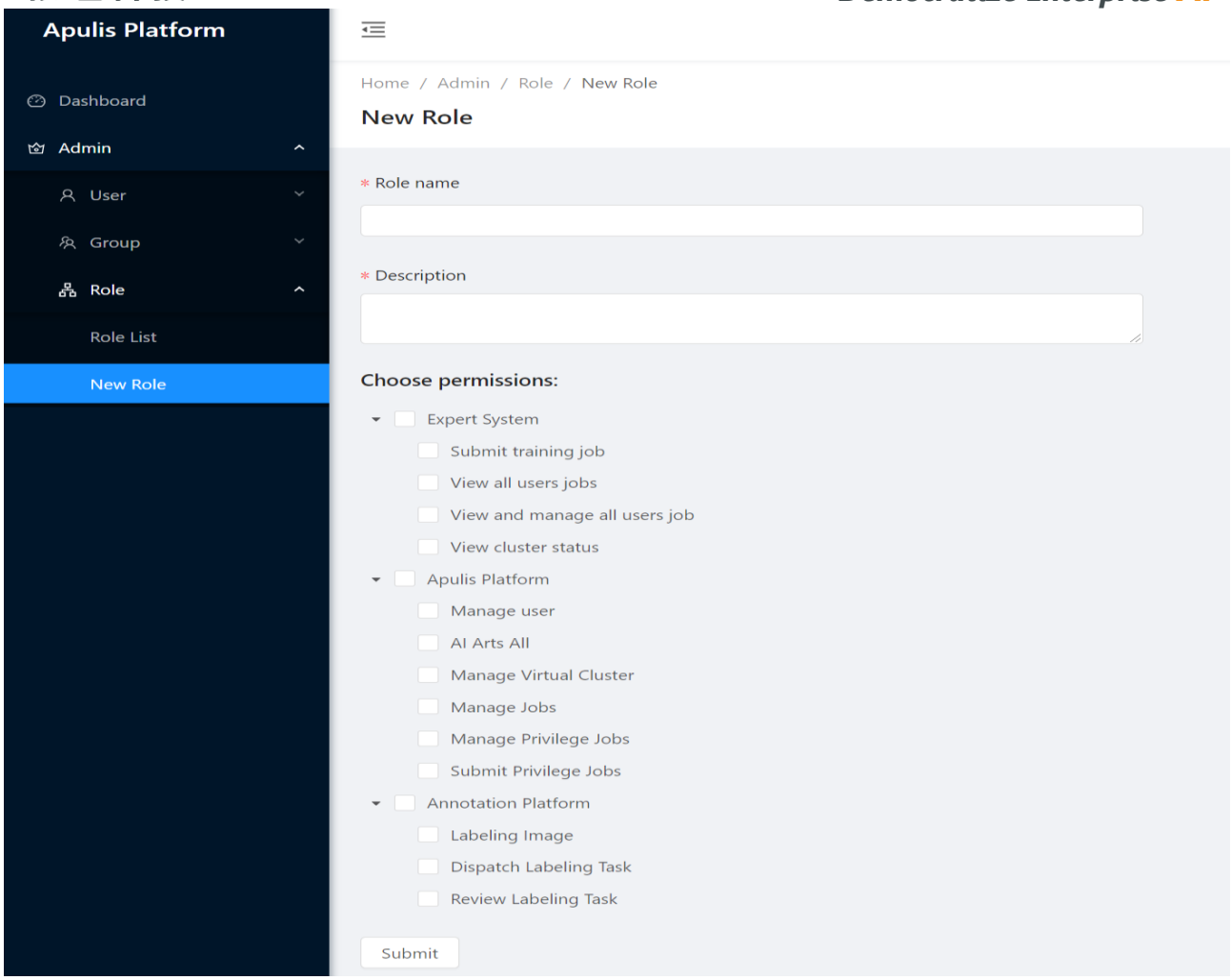


Figure 96: New role

There are two types of platform permissions:

High-performance platform: It has all permissions of the Apulis artificial intelligence platform, including the permission of the labeling platform, but not the permission of the user management system.

Annotation platform: Only has the permission to label image.

5 Expert System

Based on the original open-source platform, the expert system supports vscode plugin and Chinesse localization. We′ ve also made some improvements and optimizations. Users can concatenate a suffix `/expert` to the platform′ s URL to access the expert system, for example: http://xx.xx.xx.xx/expert

5.1 Access Expert System

Click [Expert System] in the [System] drop-down menu of the top navigation bar of Apulis AI platform.

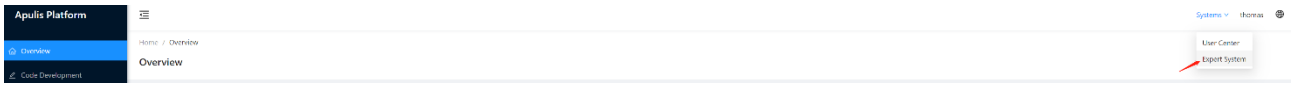


Figure 97: access expert system

5.2 View Cluster Status

Enter [Home] page to check the user′ s resource usage of VC.

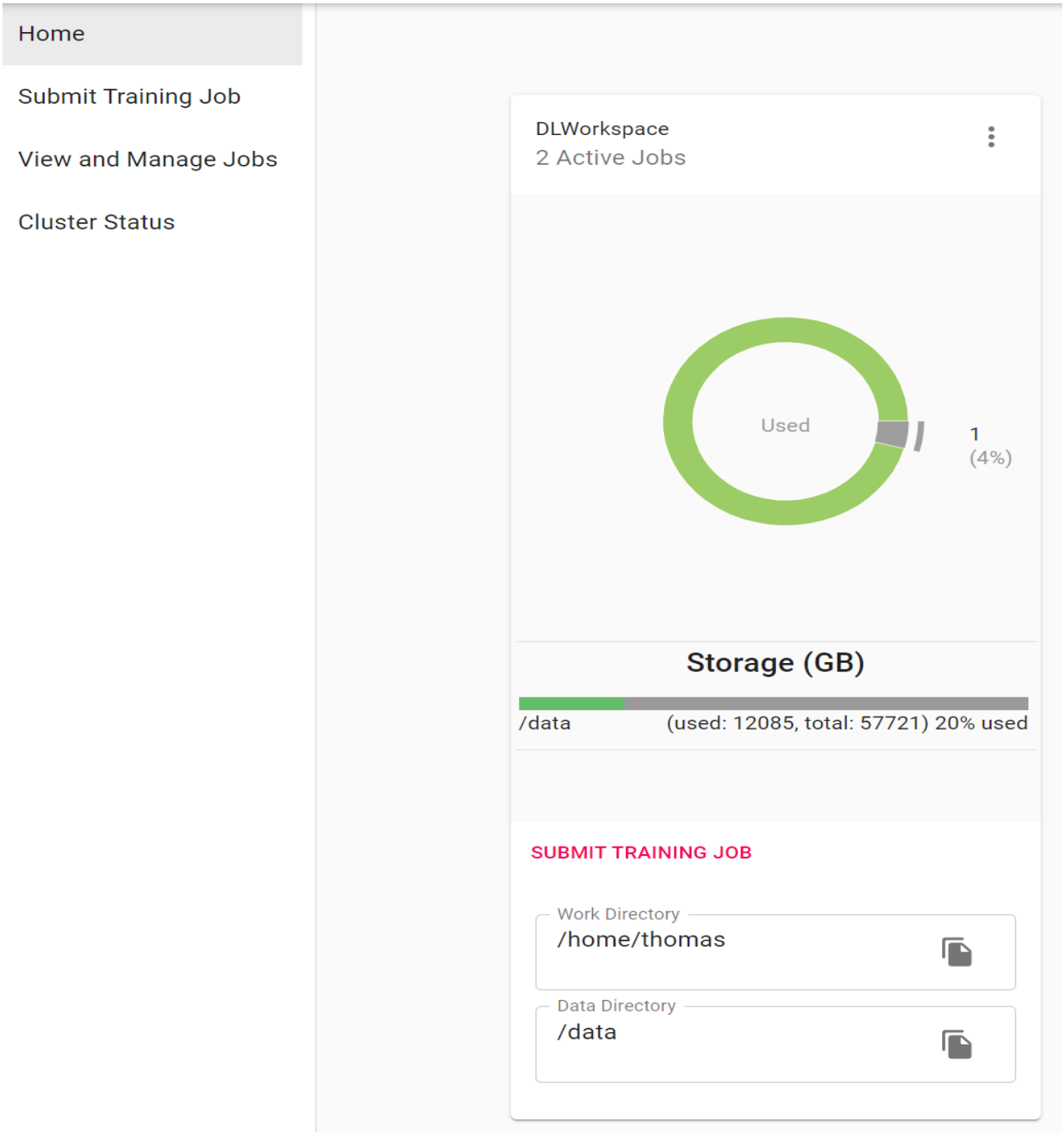


Figure 98: View cluster status

5.3 Creating Training Job

Fields for creating tasks

Name	Description

Job	Training job
Cluster	Resource cluster name
Job Type	Task type: single machine multi-card or multi-machine distributed
Device Type	Device type: x86-64 GPU, ARM64 NPU
Preemptible	Whether to allow the resource to be preempted, the default is: NO, not allowed to be preempted
Docker Image	Container, supports local mirror and docker hub public mirror and private mirror library
CMD	The executable command after starting the container, the default is bash shell
Virtual Cluster	Virtual user resource group
Edge Inference	Online edge inference, support export conversion

Please, refers to the preset template to customize the training parameters, and then click [Submit] to create a new training job or environment. Please refer to the preset template to customize the training parameters, and then click [Submit] to create a new training job or environment.

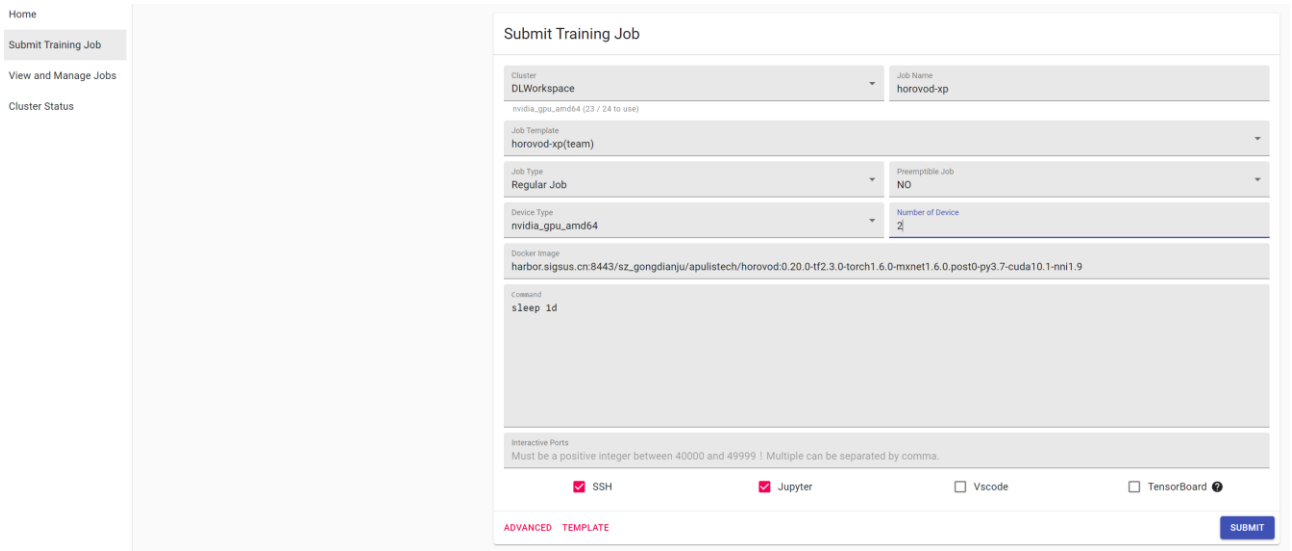


Figure 99: Create job

5.4 Create Job Template

For the convenience of colleagues, teachers and classmates, users can create training job templates according to the platform’ s environment configuration. Templates can be reused to create jobs.

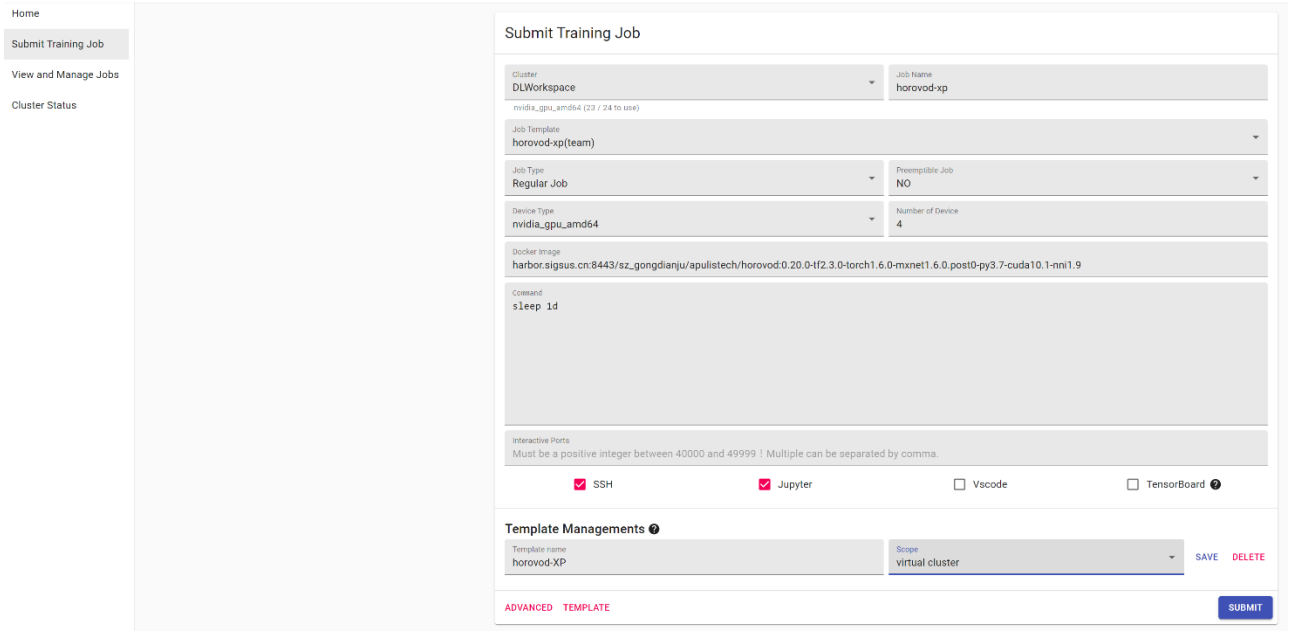


Figure 100: Create job template

5.5 Create Jobs from templates

Select the preset template in the [Job Template] drop-down menu.

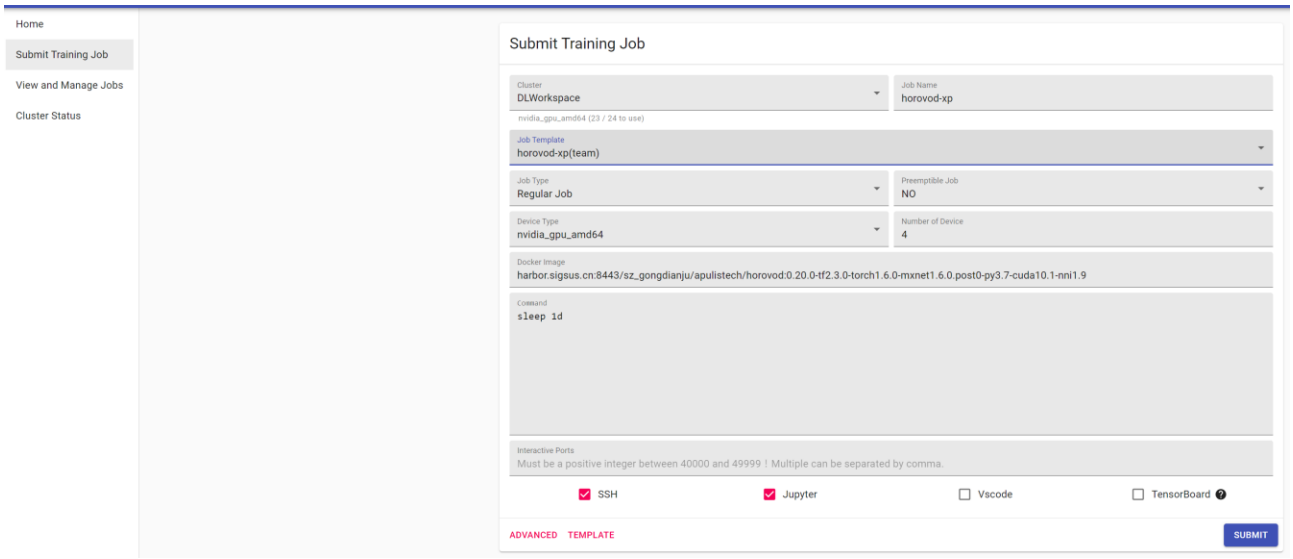


Figure 101: Create a Job from template

5.6 Job Details

After submitting the training job, users will be redirected to the [Job Detail] page, and the detailed configuration of the job can be found in the [BRIEF] tab of the [View and Manage jobs] page.

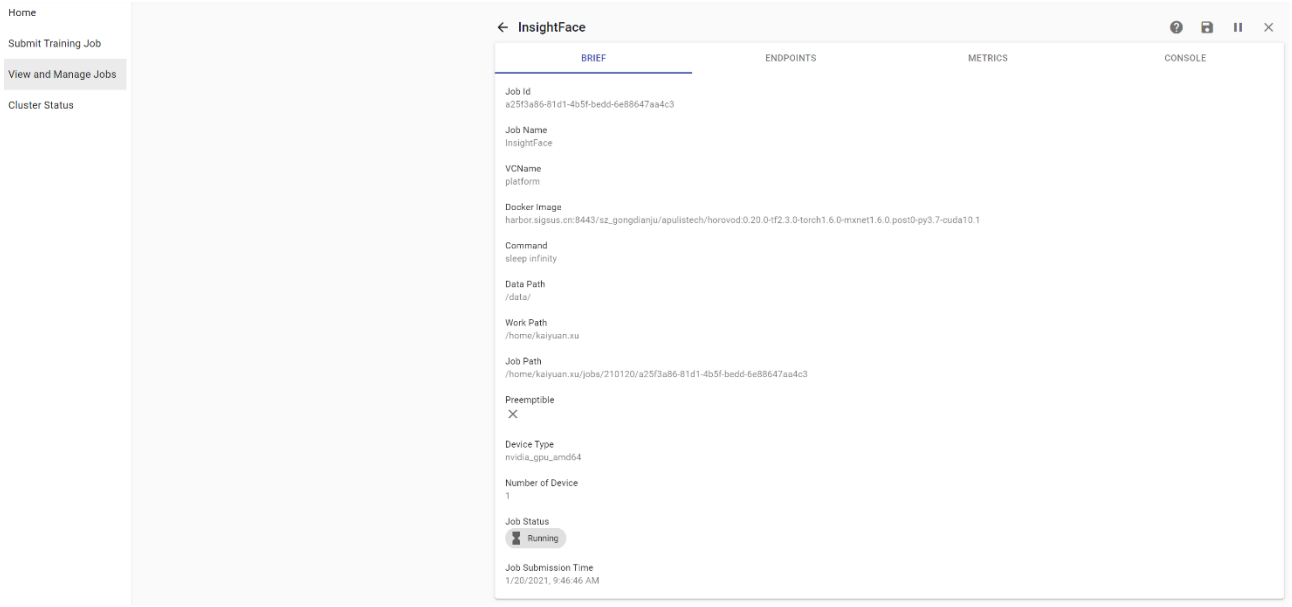


Figure 102: training job details

5.7 Manage training job status

You can start, pause or stop a job in the platform.

When you [Pause] job, the platform creates a checkpoint from which you can restart again; However, if the environment is rebuilt from a destroyed one, the platform wouldn′ t create a checkpoint for you to save the environment. But if you do want to save the environment, we recommend to create a backup or save the container image.

Note: Ordinary users can only manage jobs in the [My Jobs] tab, while administrators can manage all jobs in the [All Jobs] tab.

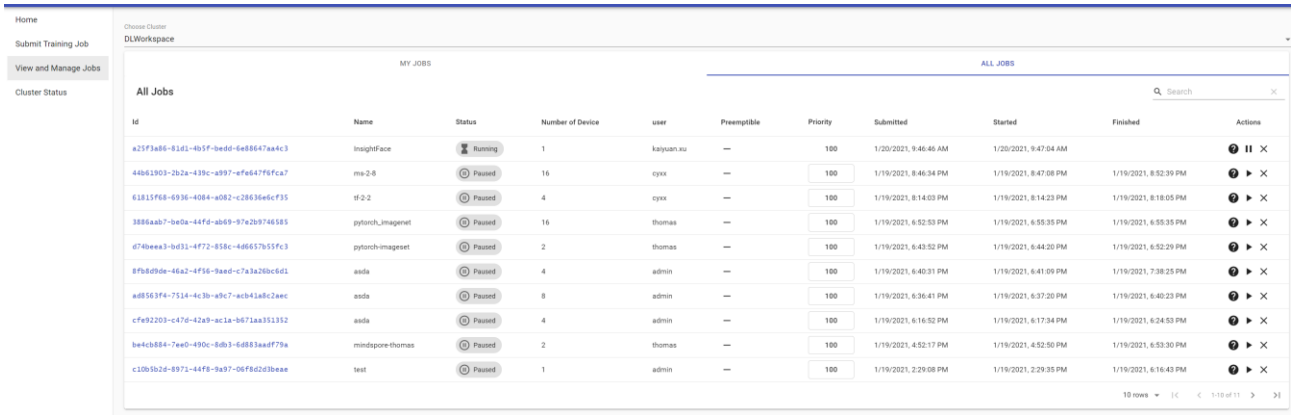


Figure 103: Manage training job status

5.8 Training Job Log

In the [CONSOLE] tab of the job detail page, users can check out the model' s training log, which includes logs of creating environment, configuring network, importing training environment variables, pulling up mirroring, running training scripts or AI framework, etc. There are tags to indicate which phase the log belongs to.

Note: The log of [Single-machine multi-card job] and [multi-node distributed job] are printed sequentially by node and card; for example, a job with 2 cards would print the log of card 1 first, and then the log of card 2. It works this way mainly because NPU driver schedules multiple cards with multiple threads. Users can also see this design in the resnet50 example.

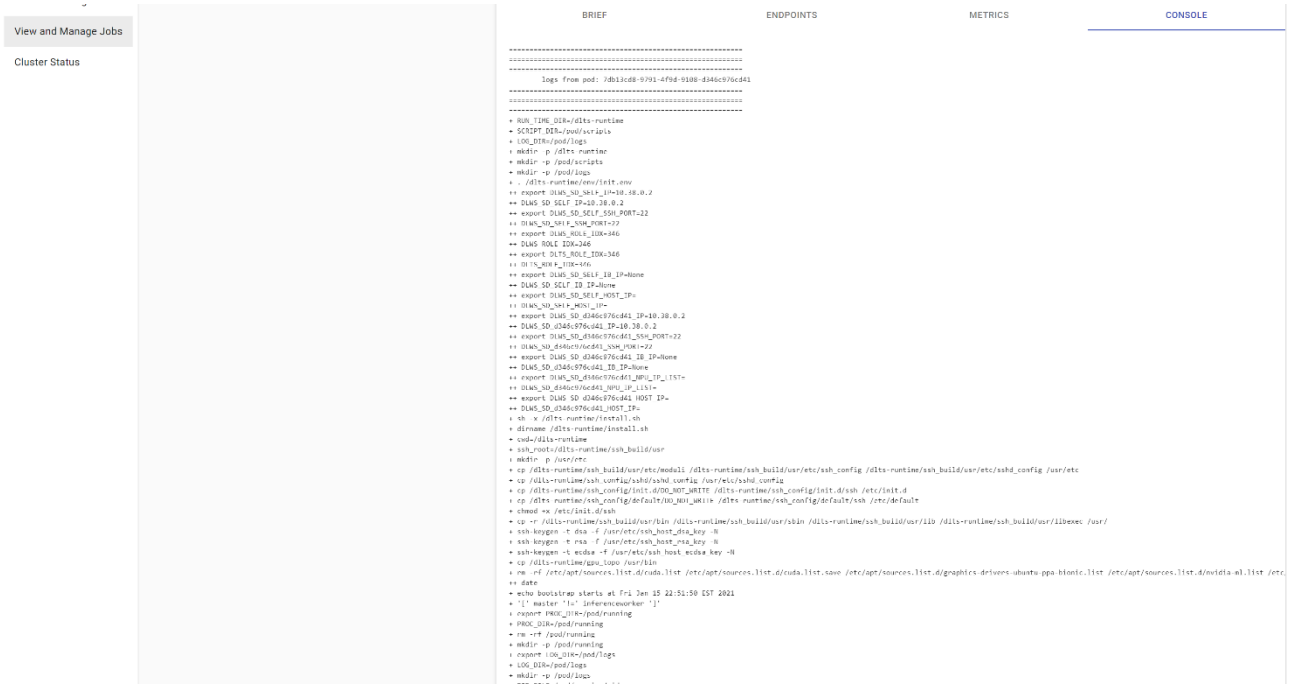


Figure 104: View training job log

The NPU-related error log is shown in the [NPU console] tab, which is extracted from the host-related error log in the `/var/log/npu/slog` directory of the host.

Since the NPU product line does not support independent export of user' s logs of the device, we export logs of all training jobs that utilize NPU. To locate problems, users could use the device ID printed in the [CONSOLE] tab during training and use it to locate error log related to this device in [NPU console].

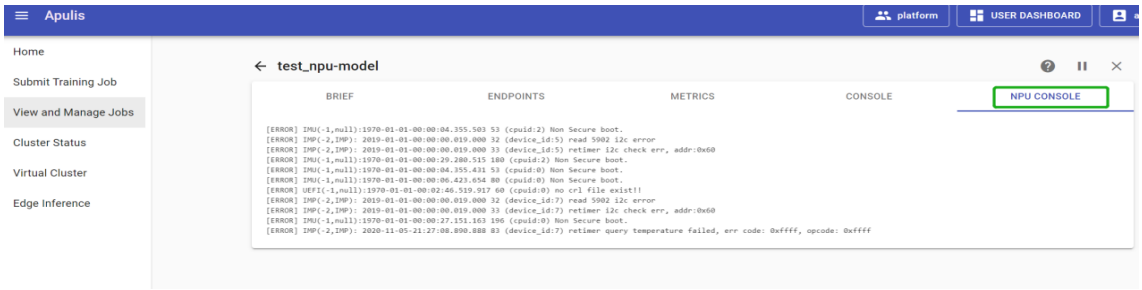


Figure 105: View NPU training job log

5.9 Resource Usage

Users can view the current job′ s resource usage in the [METRICS] of the job's details.

Figure 106: View resource usage

5.10 Interactive Debugging

In the [ENDPOINTS] tab of the job detail page, users can open interactive clients, such as Jupyter, SSH, or tensorboard.

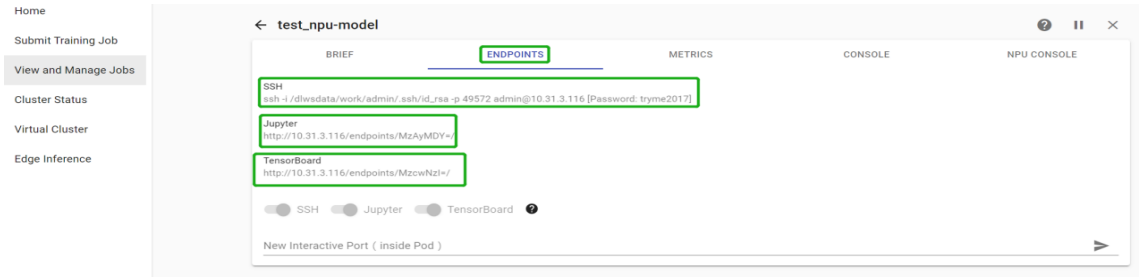


Figure 107: Interactive debugging mode

In Jupyter, ordinary users have read, write and execute permission only in the `/home` directory.

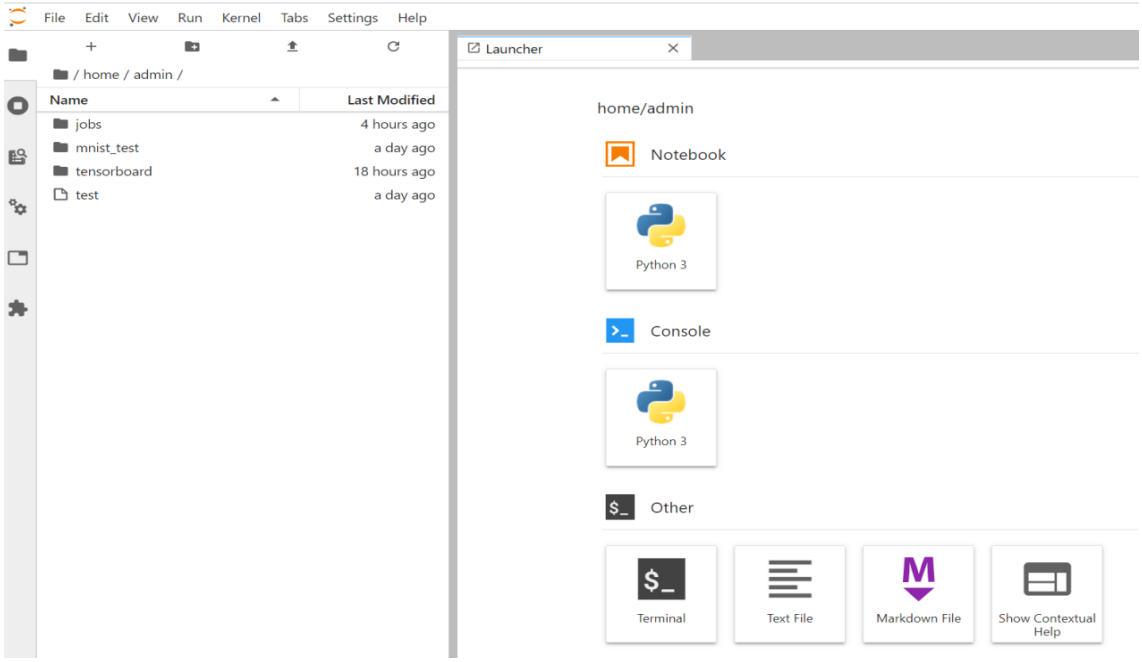


Figure 108: Jupyter debugging window

`ssh - i / dlwsdata /work/admin/. ssh / id_rsa -p 49572 admin@10.31.3.116 [Password: tryme2017]`

The link contains default RSA identity key of the host and login password; if you login through other terminals (within reachable network) to link ssh, remove the; For example: ` ssh -p 49572 admin@10.31.3.116` and enter the password: `tryme2017`

```
tomas@LAPTOP-5OPR6MGO:/mnt/c/Users/Admin$ ssh -p 38317 admin@
The authenticity of host '[10.31.3.116]:38317 ([10.31.3.116]:38317)' can't be established.
ECDSA key fingerprint is SHA256:ZT3w6RGa+3e8yonV3CropfP4gFCYOkYJ4IMY0D46Hf8.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '[10.31.3.116]:38317' (ECDSA) to the list of known hosts.
admin@10.31.3.116's password:
admin@4195bb6a-bda7-4b57-b309-e7a2f5f9b2bb:~$
```

Figure 109: terminal SSH login job environment

Tensorboard is a log visualization service. The log output path has been configured for the resnet50 preset model, but if it is a custom model, users need to configure the log path accordingly.

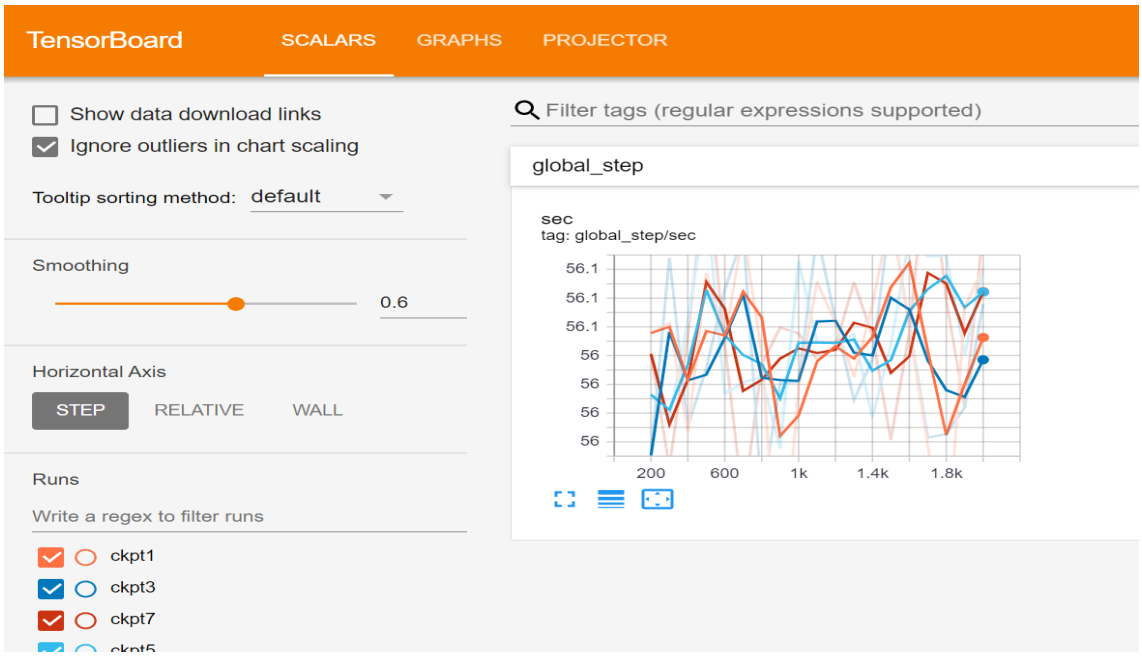


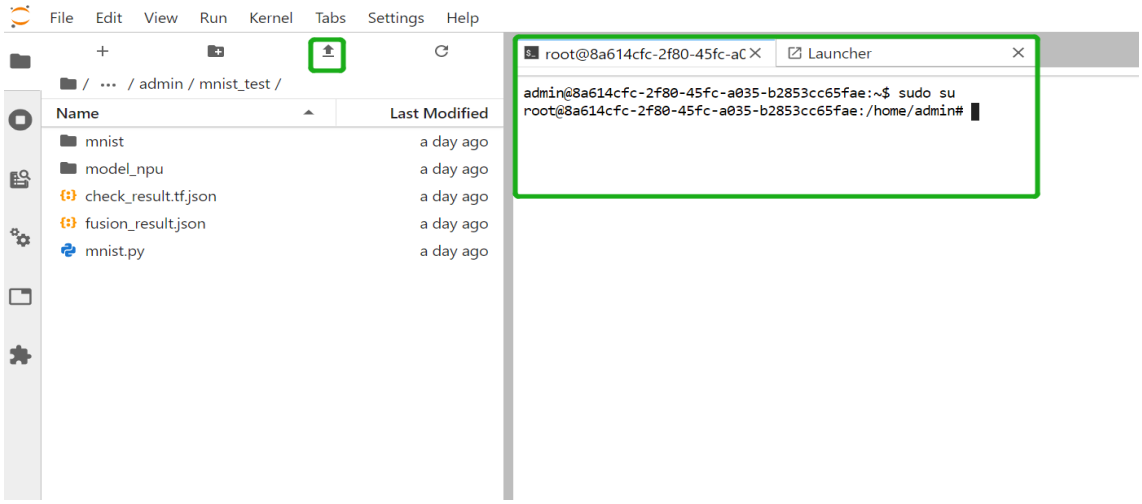
Figure 110: Use Tensorboard to analyze training logs

6 Common problem handling

6.1 Jupyter Lab interactive development

In the " /home/<USERNAME> " directory, you can upload and download files, execute scripts, or connect remote terminal.

If it is NPU-related training, you′ ll need root authority to execute npu_bridge. It is recommended to use "sudo su" to change to root environment, as shown in the figure below:

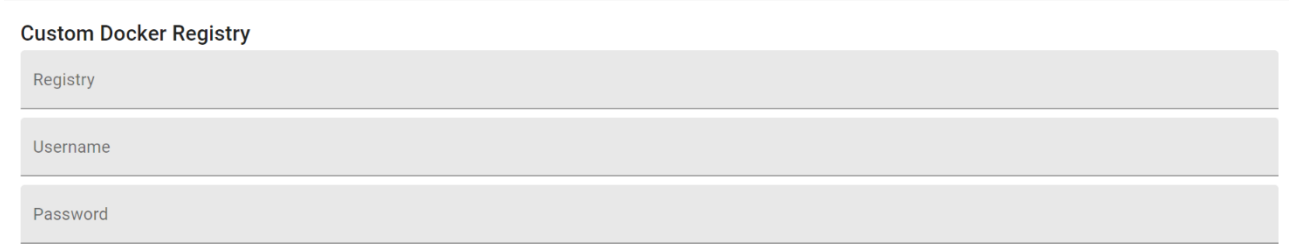


To save your trouble of environment configuration, the platform has configured commonly-used environment variables to the root and logged-in user environment. If the current user is neither a logged-in user nor a root user, you can manually load " source /pod.env ". Afterwards, the commonly-used tensorflow and npu-related environment variables will be loaded into the training environment. As shown below:

6.2 Customize Container Repository

In the [Submit Training Job] page, you can add a custom container repository in the [Advanced] options by filling in the [Custom Docker Registry] section, as shown in the following figure. Note that only https links are supported

Users can add a custom container repository with the option [Custom Docker Registry] under the advanced options [Advanced] on the [Submit Training Job] page of the submit task window. Only HTTPS links are supported. As shown below:



6.3 NPU Resource Scheduling Strategy

For NPU type of Job: the number of devices is limited to 0, 1, 2, 4, 8 for [Regular Job], while for [Distributed Job] the number of devices per node is default 8 ([Number of Nodes] should not exceed the total number of nodes). For nodes, the 1st to 3rd cards is a group, and the 4th to 7th cards is another group. The resources allocated to a job cannot distributed across groups (for example: the 3rd and 4th card are not allowed to be allocated to a single job).