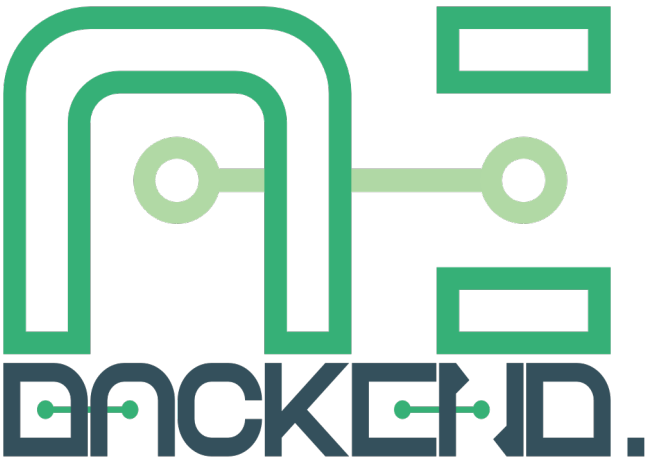


Backend.AI Console User's Guide

(19.09)

Lablup Inc.

Table of Contents

0 Disclaimer **2**

1 Overview **4**

1.1 Key Concepts 5

2 Installation **8**

3 Signup and Login **9**

3.1 Signup 9

3.2 Login 11

3.3 When you forgot your password 12

4 Summary Page **14**

4.1 Monitoring resource usage 14

4.2 Change password 16

4.3 Change SSH Keypair 17

5 Querying Compute Sessions **19**

6 Using Compute Session **20**

6.1 Start a new session 20

6.2 Use and Manage Running Session 23
6.3 Advanced Web Terminal Usage 27

7 Handing Folders **30**
7.1 Create storage folder 31
7.2 Explore folder 32
7.3 Rename folder 33
7.4 Delete folder 33
7.5 Automount folder 34

8 Create a Compute Session with Mounted Folders **35**
8.1 Configuring a compute session environment using an automount folder 37

9 Connect to a Compute Session with SSH/SFTP **40**
9.1 SSH/SFTP connection to a compute session (Linux/Mac) 40
9.2 SSH/SFTP connection to a compute session (Windows / FileZilla) . . . 43

10 Admin Menus **46**
10.1 Create and update user 47
10.2 Deactivate user account 49
10.3 Manage User's Keypairs 50
10.4 Manage Resource Policy 52
10.5 Manage Images 56
10.6 Manager docker registry 58
10.7 Manage resource preset 60
10.8 Query agent nodes 62
10.9 Manage resource group 63
10.10 System settings 64
10.11 Server management 64

11 Trouble Shooting **66**
11.1 SFTP disconnection 66

11.2 Inconsistency between allocated and actually using resources 67
11.3 Image is not displayed after it is pushed to a docker registry 67

12 Terms of License Agreement **69**
12.1 Backend.AI License (Software) 69

13 References **72**

User's guide for the Backend.AI GUI Console.

Backend.AI GUI Console is a web or app that provides easy-to-use GUI interface to work with the Backend.AI server.

The latest versions of this document can be found from sites below:

- https://console.docs.backend.ai/en/stable (English)
- https://console.docs.backend.ai/ko/stable (Korean)

CHAPTER 0

Disclaimer

The information and content in this document is provided for information and informational purposes only, and including, without limitation, any warranty of any kind, either express or implied, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose and non-infringement. AS IS. Lablup Inc. is not responsible for any damages resulting from the use of this manual, including consequential damages. Although much effort has been made to ensure the accuracy of the data provided, Lablup Inc. makes no representations, promises, or warranties with respect to the completeness, accuracy, up-to-date or adequacy of the information, and assumes no responsibility for outdated information or errors. Lablup Inc. makes no representations or warranties, either expressed or implied, with respect to the accuracy or completeness of the information contained in this manual.

Please contact us if you believe there are errors in this document that are different from the facts. We will review it as soon as possible.

The product in this manual may subect to change without notice due to continued

development by open source community and Lablup Inc.

All product and company names mentioned herein may be trademarks of their respective owners.

Backend.AI is a registered trademark of Lablup Inc.

The copyright for Backend.AI™ is owned by Lablup Inc.

©2015-2020 Lablup Inc. All rights reserved.

Document Version: v19.09.200912

Last Updated (Year/Month/Day): 2020/09/12

CHAPTER 1

Overview

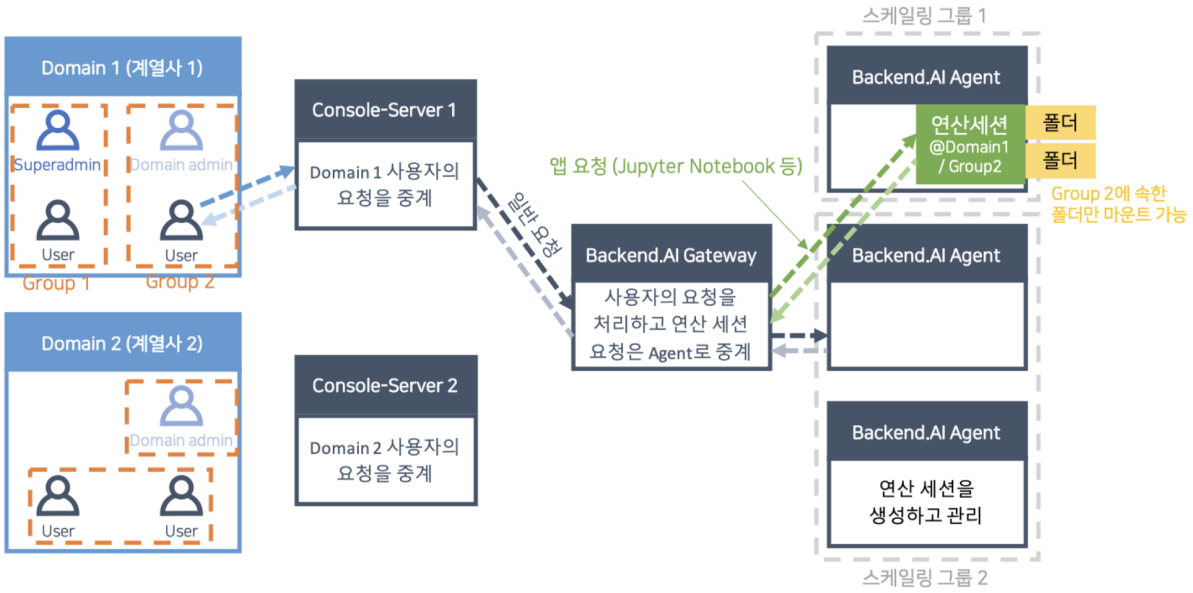
Backend.AI is an open source cloud resource management platform. Backend.AI efficiently manages compute resources in a cloud or on-premises cluster, providing a virtualized compute environment to provide on-demand computation, anytime, anywhere. With GPU virtualization technique, Backend.AI helps scientists, DevOps, enterprise, and AI enthusiasts to scale up efficiently.

Backend.AI offers a variety of performance-driven optimizations for machine learning and high-performance computing clusters, along with management and research features to support a diverse users, including researchers, administrators, and DevOps. The Enterprise Edition adds support for multi-domain management, a dedicated Hub app for superadmins, and the GPU virtualization plug-in.

A GUI client package is also provided to easily take advantage of the features supported by the Backend.AI server. Backend.AI Console is a GUI client in the form of a web service or standlone app. It provides a convenient graphical interface for accessing the Backend.AI server to utilize computing resources and manage its environment. Most

tasks can be done with mouse clicks and brief typing, which achieves more intuitive use.

1.1 Key Concepts



- User: The user is the person who connects to Backend.AI and performs work. Users are divided into normal users, domain admins, and superadmins according to their privileges. While ordinary users can only perform tasks related to their computing sessions, domain admins have the authority to perform tasks within a domain, and superadmins perform almost all tasks throughout the system. A user belongs to one domain and can belong to multiple groups within a domain.
- Compute session, container: An isolated virtual environment in which your code runs. It looks like a real Linux server with full user rights, and you can't see other user's session even if it's running on the same server as your session. Backend.AI implements this virtual environment through a technology called containers. You can only create compute sessions within the domain and groups to which you belong.
- Domain: This is the top layer for authority and resource control supported by

Backend.AI. For companies or organizations, you can view domains as an affiliate and set up per-domain (or per-affiliate) permissions and resource policies. A user should belong to only one domain, and can create sessions or do some other jobs only in their own domain. A domain can have one or more domain admins, who can set policies within the domain or manage sessions. For example, if you set the total amount of resources available in a domain, the resources of all containers created by users in the domain cannot be greater than the amount set.

- Groups: A hierarchy belonging to a domain. Multiple groups can exist in one domain. You can think of a group as a project unit. A user can belong to multiple groups (or projects) at the same time within a domain. Compute sessions must belong to one group, and users can only create sessions within their own groups. Domain admins can set policies or manage sessions for groups within the domain. For example, if you set the total amount of resources available within a group, the resources in all containers created by users in the group cannot be greater than the amount set.

- Image: Each container has a pre-installed language-specific runtime and various computational frameworks. The state of such snapshots before they are executed is called an image. You can choose to run an image provided by the cluster admin, create your own image with the software you want to use, and ask the admin to register it.

- Virtual Folder (vfolder): A "cloud" folder that is always accessible and mountable in a container on a per-user basis, regardless of which node the container runs on. After creating your own virtual folder, you can upload your own program code, data, etc. in advance, and mount it when you run the compute session to read from and write to it as if it is on your local disk.

- Application service, service port: A feature that allows you to access various user applications (eg DIGITS, Jupyter Notebook, shell terminal, TensorBoard, etc.) running within the compute session. You do not need to know the container's

address and port number directly, but you can use the provided CLI client or GUI console to directly access the desired daemon of the session.

- Console App: A GUI client that is served as a web or standalond app. You can use the service after logging in by specifying the address of the Backend.AI server and entering the user account information.
- Local wsproxy: Proxy server built into the console app. Services such as Jupyter Notebook and Terminal that can be used in the console app communicate with the server through websocket, which converts general HTTP requests to and from the console app into websocket to deliver messages.
  - If the console app loses its connection to wsproxy or the wsproxy server is dead, it will not be possible to access services such as Jupyter Notebook and Terminal.
- Web wsproxy: In the case of the console app provided as a web, the built-in server cannot be operated due to the nature of the browser. In this case, you can use services such as Jupyter Notebook, Terminal, etc. in the web environment by making the wsproxy server as a separate web server so that the console app can see the web wsproxy.

CHAPTER 2

Installation

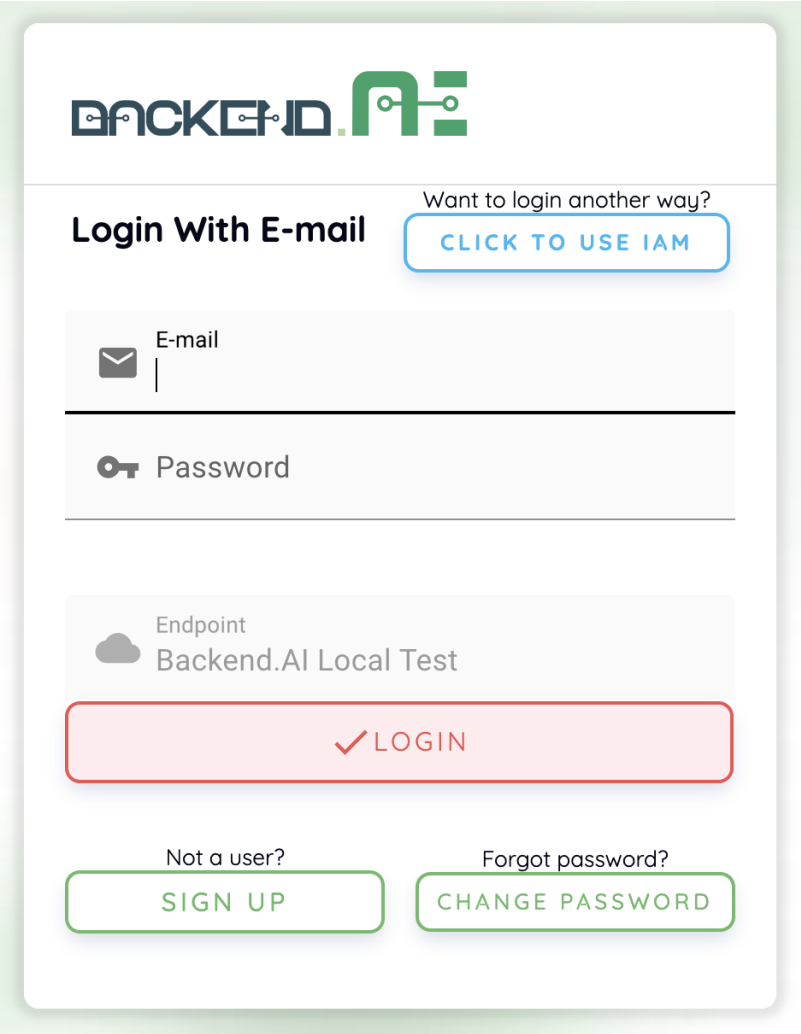
The GUI console can be used in two forms. It can be used as a web service by connecting to a separate web address prepared by the admins, or as an app provided as a standalone executable that does not require separate installation. In the case of an app, depending on the security setting of the desktop OS, it may be recognized as an unsigned executable file and a permission check may be required.

CHAPTER 3

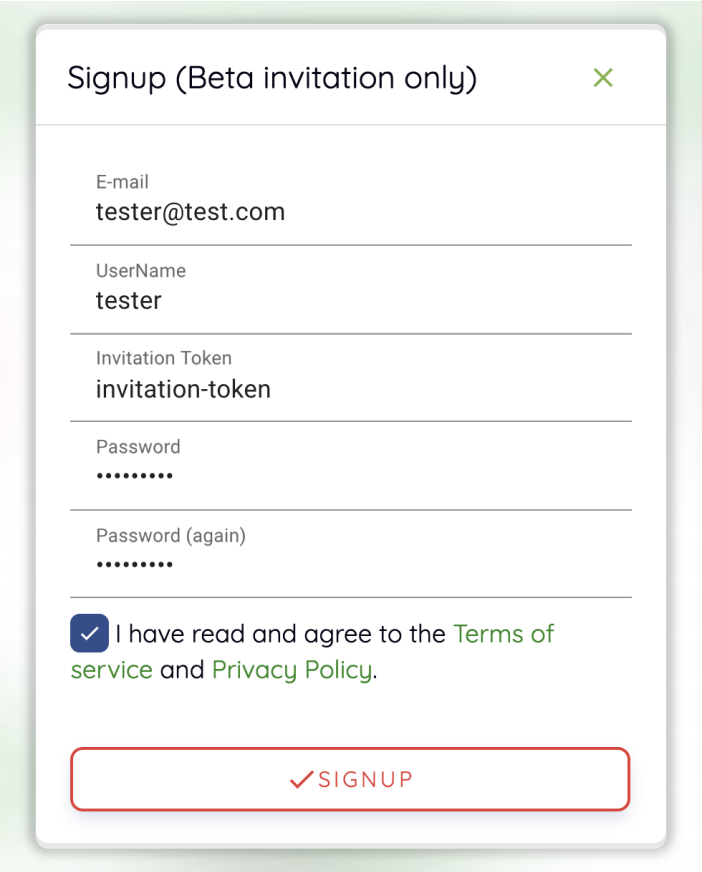
Signup and Login

3.1 Signup

When you launch the GUI console, login dialog appears. If you haven't signed up yet, press the SIGN UP button.



Enter email, username, password, etc., read and agree to the Terms of Service / Privacy Policy, and click the SIGNUP button. Depending on your system settings, you may need to enter an invitation token to sign up. Furthermore, an email may be sent to verify that the email is yours. If a verification email is sent, you will need to read the email and click the link inside to pass verification before you can log in with your account.



Note: Depending on the server configuration and plugin settings, sign up by anonymous user may not be allowed. In that case, please contact administrator of your system.

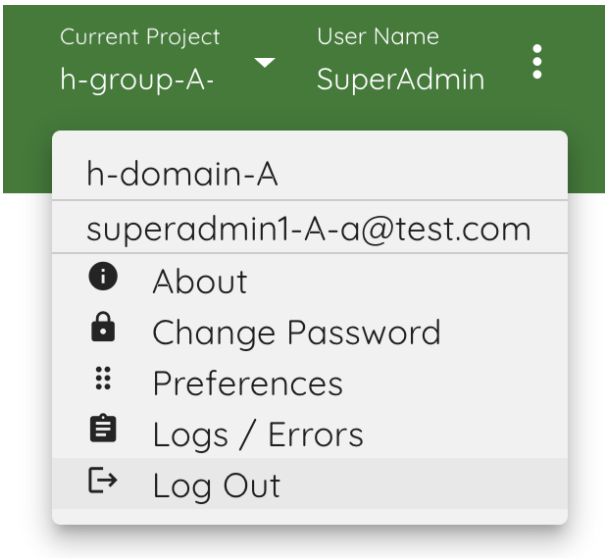
3.2 Login

Enter your ID and password and press the LOGIN button. In API ENDPOINT, the URL of Backend.AI Console Server, which relays the request to the Manager, should be entered.

Note: Depending on the installation and setup environment of the Console Server, the endpoint might be pinned and not configurable.

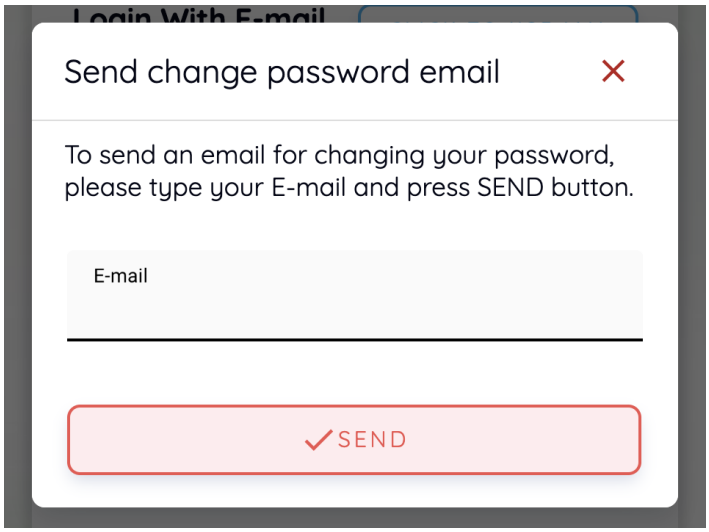
After login, you can check the information of the current resource usage in the Summary tab.

Buy clicking the icon in the upper-right corner, you will see sub menus. You can logout by selecting the Log Out menu.



3.3 When you forgot your password

If you have forgotten your password, you can click the CHANGE PASSWORD button on the login panel to email a link to change password. You can change your password by reading and following the instruction. Depending on the server settings, the password change feature may be disabled. In this case, contact the administrator.



Note: This is also a modular feature, so changing password may not be possible in some systems.

CHAPTER 4

Summary Page

The Summary page allows you to view resource and session usage information. There are also shortcut links to frequently used features.

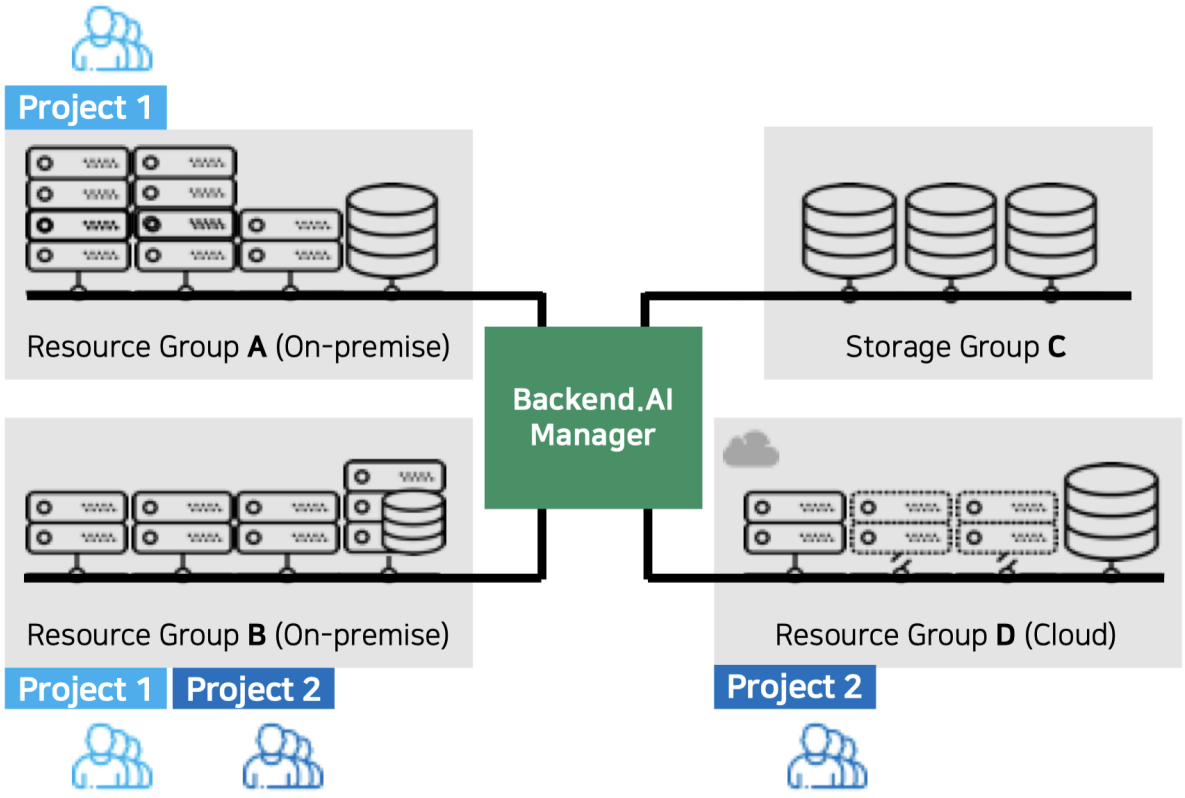
4.1 Monitoring resource usage

In the Start menu, you can see the total amount of resources you can allocate and how much you are currently using. You can check the CPU, RAM, GPU occupancy status, the number of sessions currently running, etc. You can also click the + START button to create a session immediately (see the Sessions page for how to create a session).

A user can belong to one or more groups. If you want to change the group you want to use, click the Project button on the top right corner to select the group you want. When the group is changed, the total amount of available resources is changed according to the resource usage policy set in the group, and the amount of resources currently used is updated with the amount of resources occupied by sessions running in the group. The

group to which the user belongs is set up by the administrator. If you want to change the group, you should contact the administrator in charge.

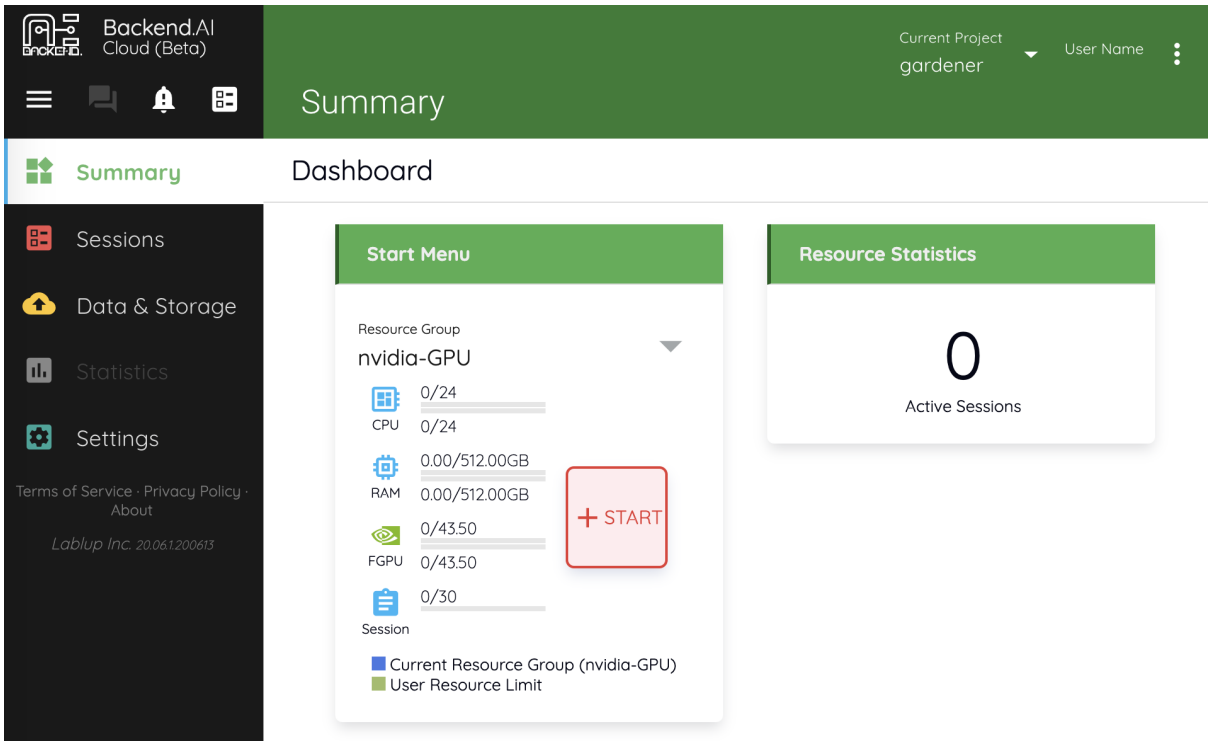
A user can be set up to use multiple resource groups by an administrator. Resource groups are administrator functions that allow you to set up worker nodes that can be accessed by user, group, or domain. For example, if you group multiple V100 GPU servers into one resource group, group multiple P100 GPU servers into another resource group, and make specific users accessible only to the P100 resource group, that user Only resources on the server are available, not resources on the V100 GPU server. You can select a resource group to which you have access by clicking Resource Group within the Start Menu. Each time a resource group is changed, the amount of available resources is changed accordingly. Resource group policies are set up by administrators. If you want to use a particular resource, you should contact your administrator.



In the Project section at the bottom, you can see the total usage within your group and your usage within that group.

The Resource Statistics panel shows the total number of sessions currently running by

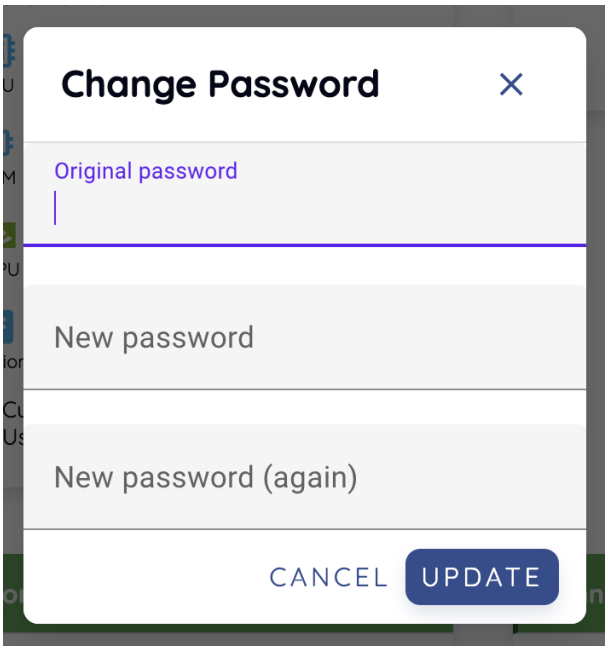
the user. Other session related information will be displayed depending on the user's permission.



4.2 Change password

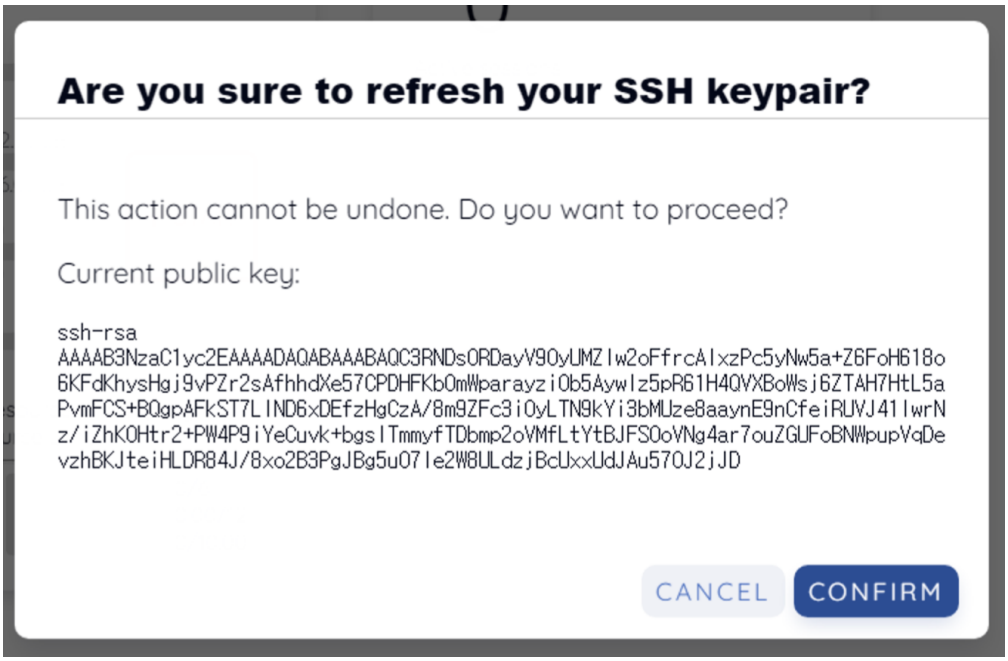
You can change the password by clicking the Change Password menu of the settings button in the upper right corner. Enter the current password in the Original password field and the new password in the New password and New password (again) fields, then press the UPDATE button to update the password. At this time, the password should follow the rules below:

- 8 letters or more
- One or more alphabets
- One or more special characters
- One or more numbers



4.3 Change SSH Keypair

If you are using the GUI console app, you can create SSH/SFTP connection directly to the container. You can query or create a SSH keypair, which is used for SSH/SFTP connection. Clicking the Refresh SSH Keypair link in the Preferences panel brings up a dialog. Current public key is displayed, if exists, and new public key and private key can be created by clicking CONFIRM button. This will generate a new SSH keypair and save it into the database.



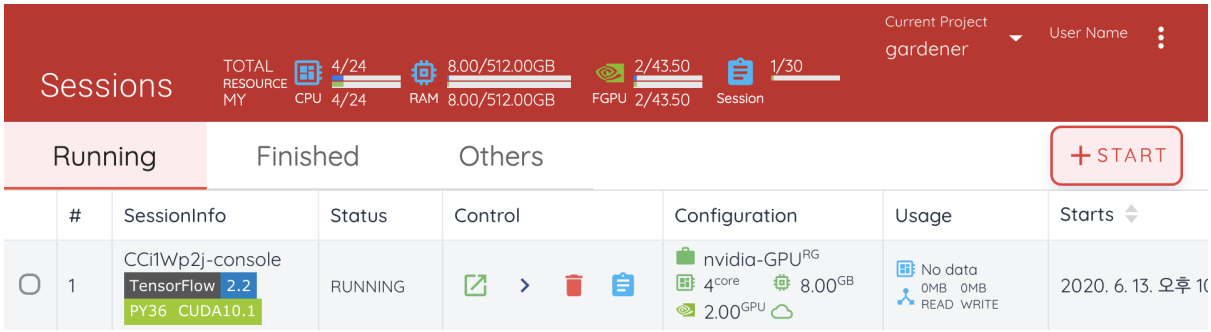
Note: The web-based console does not yet support SSH/SFTP connections.

Note: Backend.AI uses SSH keypair based on OpenSSH. On Windows, you may convert this into PPK key.

CHAPTER 5

Querying Compute Sessions

To see the list of compute sessions, click Sessions in the left sidebar. In the Running tab on the right, you can check the information on the currently running sessions. Click the Finished tab to see the list of terminated sessions. For each session, you can check information such as ID, created date, used time, allocated resources, resource usage, and etc.



As a superadmin, you can see the information of all sessions currently running (or ended) in the cluster. On the other hand, users can see their sessions only.

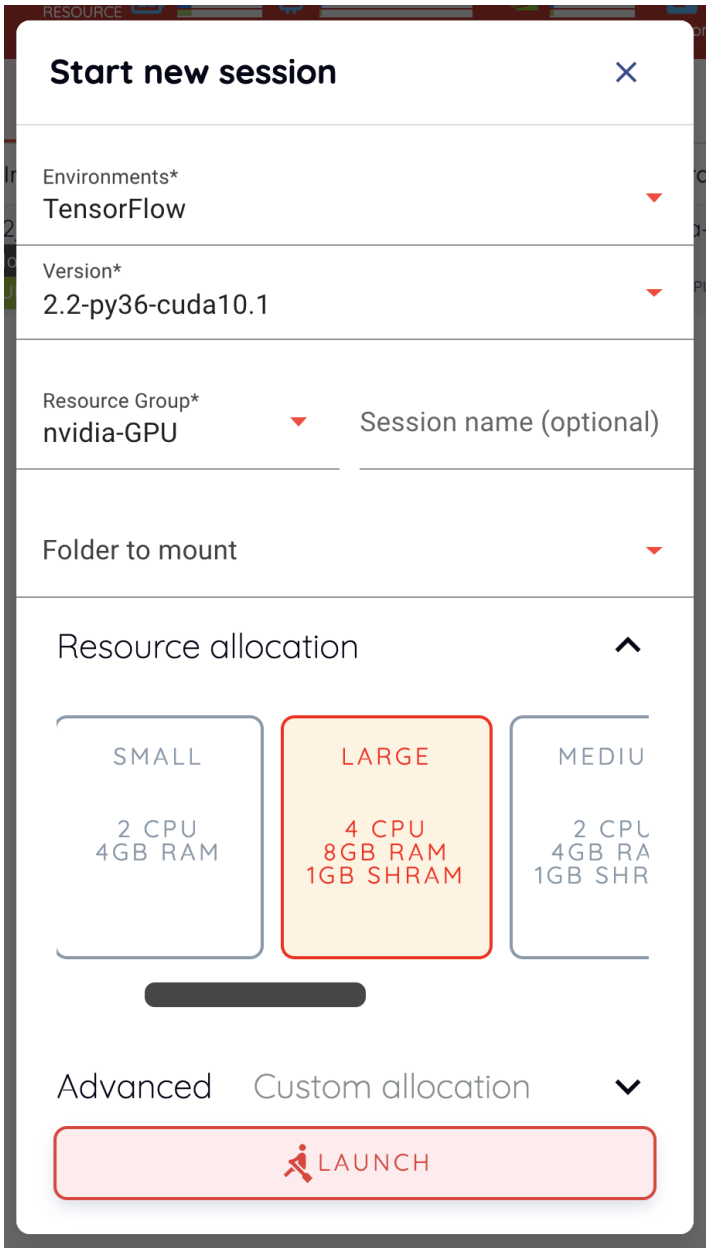
CHAPTER 6

Using Compute Session

In addition to see the list of compute sessions, Sessions tab lets you start new sessions or use and manage already running sessions.

6.1 Start a new session

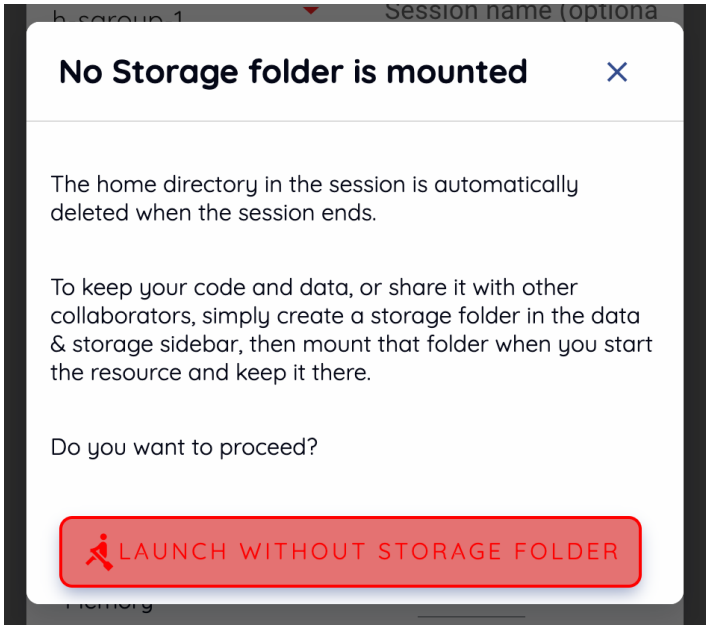
Click START button to start a new compute session. The following setup dialog will appear. Specify the language environment (Environments, Version), the amount of resources (CPU, RAM, GPU, etc.) you want to use, and then press the LAUNCH button.



Note: If the GPU resource is marked as FGPU, this means that the server is serving the GPU resources in a virtualized form. Backend.AI supports GPU virtualization technology that a single physical GPU can be divided and shared by multiple users for better utilization. Therefore, if you want to execute a task that does not require a large amount of GPU computation, you can create a compute session by allocating only a portion of a GPU. The amount of GPU resources that 1 FGPU actually allocates may vary from system to system depending on the administrator's setting. For example, if administrator has set to split one physical GPU into five pieces, 5 FGPU means 1 physical GPU, or 1

FGPU means 0.2 physical GPU. At this configuration, if you create a compute session by allocating 1 FGPU, you can utilize SM (streaming multiprocessor) and GPU memory corresponding to 0.2 physical GPU for the session.

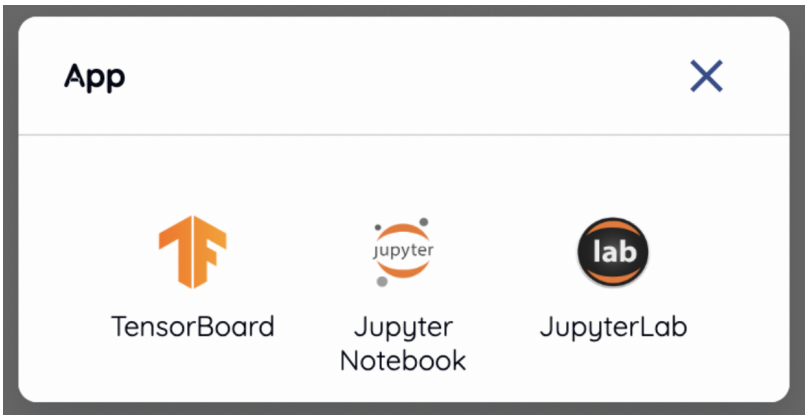
If no mount folder is specified in "Folders to mount", the following notification dialog may appear. It is recommended that one or more storage folders to be mounted because terminating compute session by default deletes all the data inside the session. If you specify a mount folder and save your data in that folder, you can keep the data even if the compute session is destroyed. Data preserved in the storage folder can also be reused by re-mounting it when creating another compute session. You can ignore the alarm and create a session. However, it's a good idea to mount a folder if you're working on a job that requires you to keep data from within a session. For information on how to mount a folder and run a compute session, see *Related Content*.



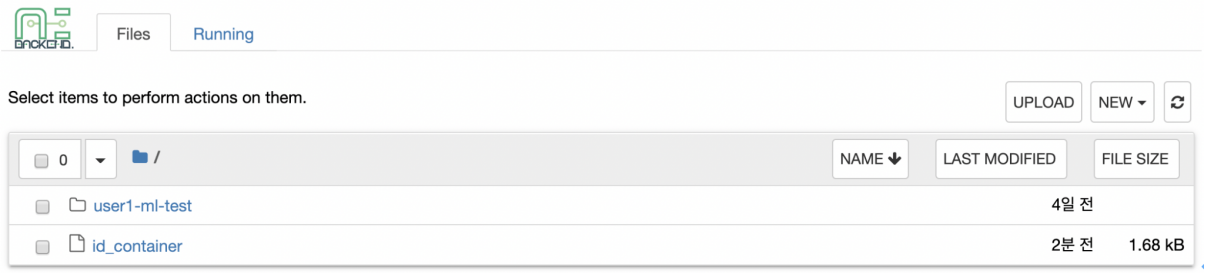
Notice that a new compute session is created in the Running tab.

6.2 Use and Manage Running Session

This time, let's take a look at how to use and manage a running compute session. If you see the Control column in the session list, there are several icons. When you click the first icon, several app services supported by the session will appear as shown in the following figure.



As a test, let's click on Jupyter Notebook.

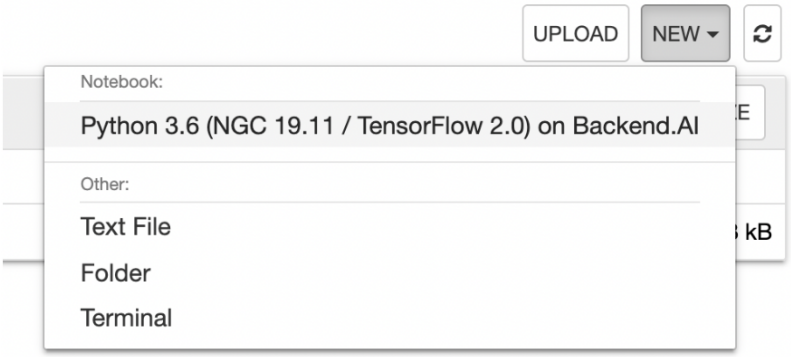


You will see a new window pop up and Jupyter Notebook is running. This Notebook was created inside the running compute session, and it's easy to use with just a click of

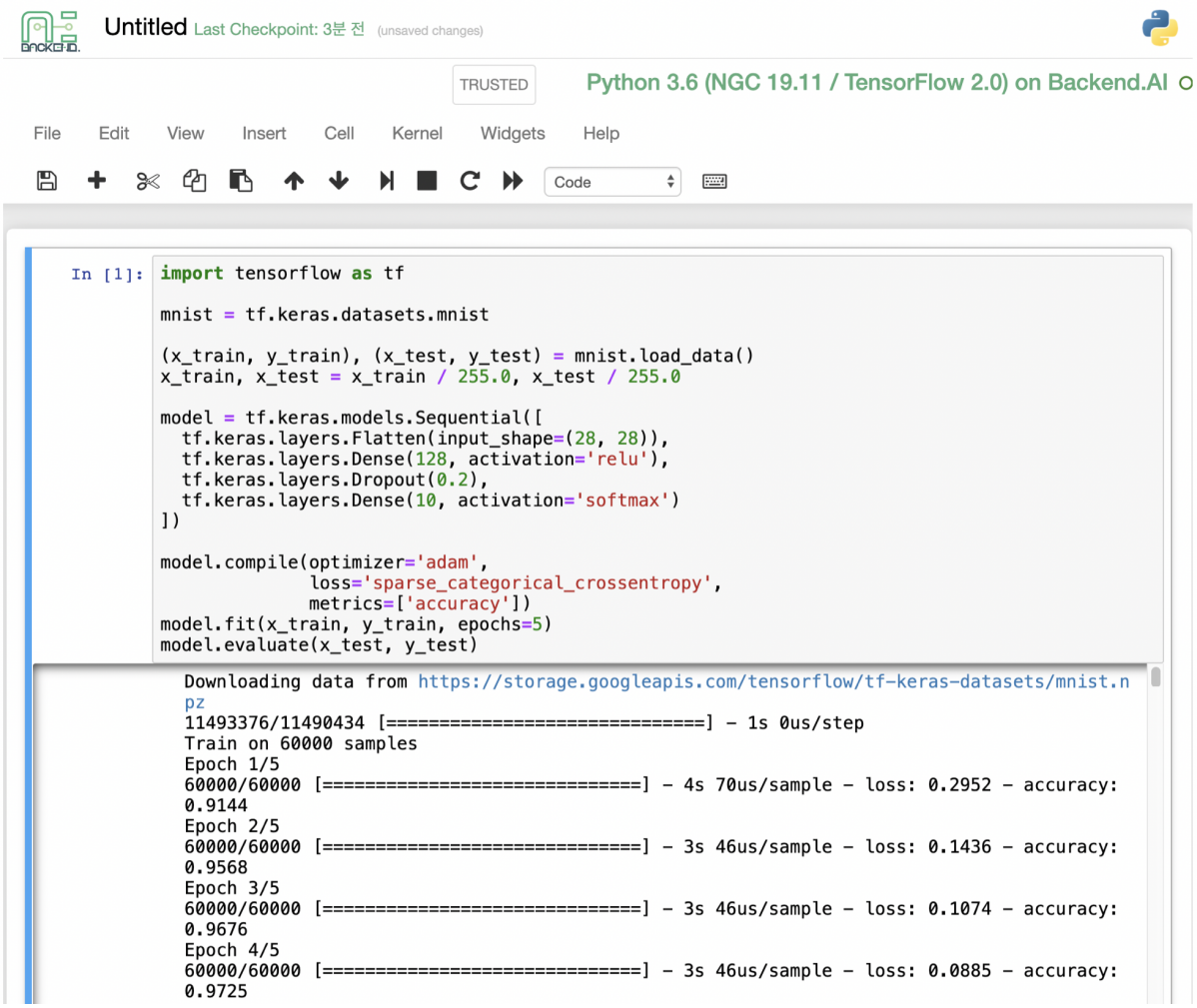
a button without any setup. In addition, you can just use the language environment and libraries provided by the compute session as is, so there is no need to install a separate packages. For more information on how to use Jupyter Notebook, please refer to the official documentation.

In the notebook's file explorer, the `id_container file` contains a private SSH key. If necessary, you can download it and use it for SSH / SFTP access to the container.

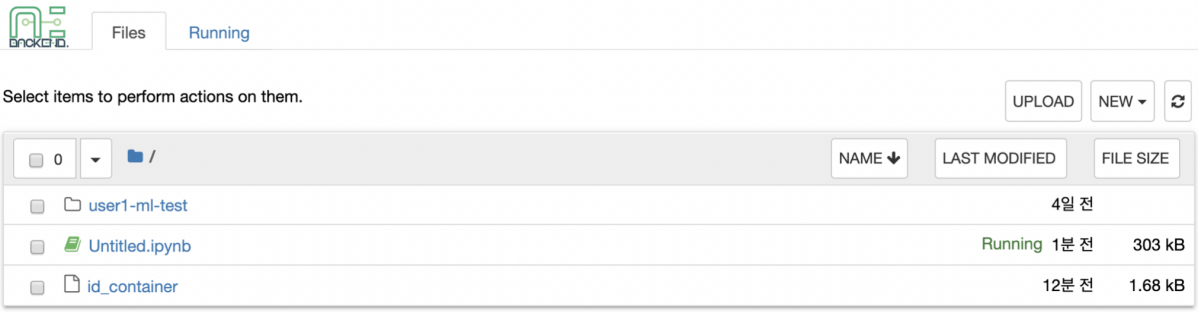
Click the NEW button on the upper right corner and select Notebook for Backend.AI, and ipynb window will pop up where you can enter the new code.



In this window, you can enter and execute any code you want by using the environment that session provides. The code execution happens on one of the Backend.AI nodes where the compute session is actually created, and there is no need to configure a separate environment on the local machine.

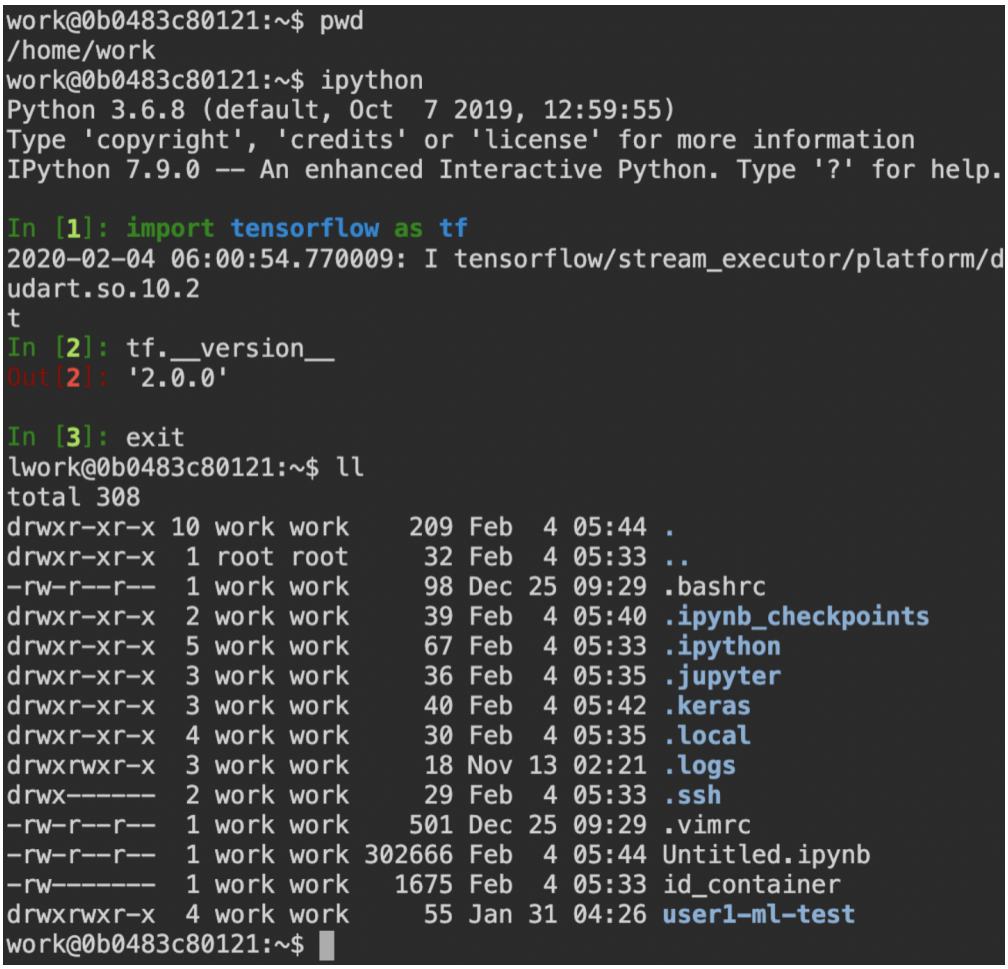


When you close the window, you can notice that the `Untitled.ipynb` file is created in the Notebook File Explorer. Note that the files created here are deleted when you destroy the session. The way to preserve those files even when the session is gone is described in the Storage/Folders section.



Return to the Session list page. This time, let's launch the terminal. Click the terminal icon (the second button) to use the container's ttyd daemon. The terminal will also appear in a new window, and you can type commands, just like any usual terminal,

which will be delivered to the compute session as shown in the following figure. If you are familiar with using command-line interface (CLI), you can easily interact with Linux commands.



If you create a file here, you can immediately see it in the Jupyter Notebook you opened earlier as well. Conversely, changes made to files in Jupyter Notebook can also be checked right from the terminal. This is because they are using the same files in the same compute session.

In addition, you can use web-based services such as TensorBoard, Jupyter Lab, etc., depending on the type of services provided by the compute session.

To delete a specific session, tap the red trash icon. Since the data in the folder inside the compute session is deleted as soon as the compute session ends, it is recommended that you move the data to the mounted folder or upload it to the folder from the beginning if you want to keep it.

6.3 Advanced Web Terminal Usage

The web-based terminal we used above internally embed a utility called tmux. tmux is a terminal multiplexer that supports to open multiple shell windows within a single shell, so as to allow multiple programs to run in foreground simultaneously. If you want to take advantage of more powerful tmux features, you can refer to the official tmux documentation and other usage examples on the Internet.

Here we are introducing some simple but useful features.

Copy terminal contents

tmux offers a number of useful features, but it's a bit confusing for first-time users. In particular, tmux has its own clipboard buffer, so when copying the contents of the terminal, you can suffer from the fact that it can be pasted only within tmux by default. Furthermore, it is difficult to expose user system's clipboard to tmux inside web browser, so when using tmux, the terminal contents cannot be copied and pasted to other programs of user's computer. The so-called `Ctrl-C` / `Ctrl-V` is not working.

If you need to copy and paste the terminal contents to your system's clipboard, you can temporarily turn off tmux's mouse support. First, press `Ctrl-B` key to enter tmux control mode. Then type `:set -g mouse off` and press `Enter` (note that you have to type the first colon as well). You can check what you are typing in the status bar at the bottom of the screen. Then drag the desired text from the terminal with the mouse and press the `Ctrl-C` or `Cmd-C` (in Mac) to copy them to the clipboard of the user's computer.

With mouse support turned off, you cannot scroll through the mouse wheel to see the contents of the previous page from the terminal. In this case, you can turn on mouse support again. Press `Ctrl-B`, and this time, type `:set -g mouse on`. Now you can scroll mouse wheel to see the contents of the previous page.

If you remember `:set -g mouse off` or `:set -g mouse on` after `Ctrl-B`, you can use the web terminal more conveniently.

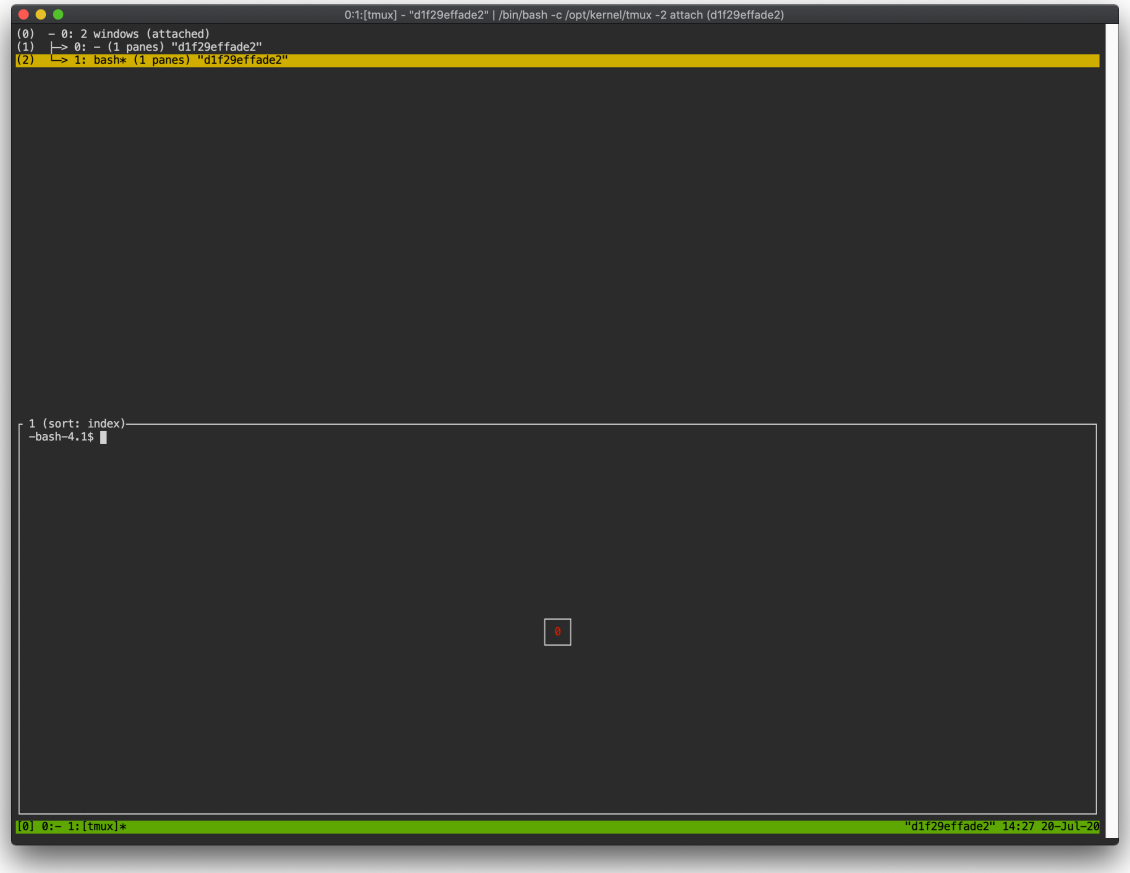
Note: `Ctrl-B` is tmux's default control mode key. If you set another control key by modifying `.tmux.conf` in user home directory, you should press the set key combination instead of `Ctrl-B`.

Checking the terminal history using keyboard

There is also a way to copy the terminal contents and check the previous contents of the terminal simultaneously. It is to check the previous contents using the keyboard. Again, click `Ctrl-B` first, and then press the `Page Up` and/or `Page Down` keys. You can see that you navigate through the terminal's history with just keyboard. To exit search mode, just press the `q` key. With this method, you can check the contents of the terminal history even when the mouse support is turned off to allow copy and paste.

Spawn multiple shells

The main advantage of tmux is that you can launch and use multiple shells in one terminal window. Since seeing is believing, let's press the `Ctrl-B` key and then the `c`. You can see that the contents of the existing window disappeared and a new shell environment appeared. Then, did the previous window terminated? It's not like that. Let's press `Ctrl-B` and then `w`. You can now see the list of shells currently open on tmux like following image. Here, the shell starting with `0:` is the shell environment you first saw, and the shell starting with `1:` is the one you just created. You can move between shells using the up/down keys. Place the cursor on the shell `0:` and press the Enter key to select it.



You can see the shell environment you saw first appears. In this way, you can use multiple shell environments within a web terminal. To exit or terminate the current shell, just enter `exit` command or press `Ctrl-B x` key and then type `y`.

In summary:

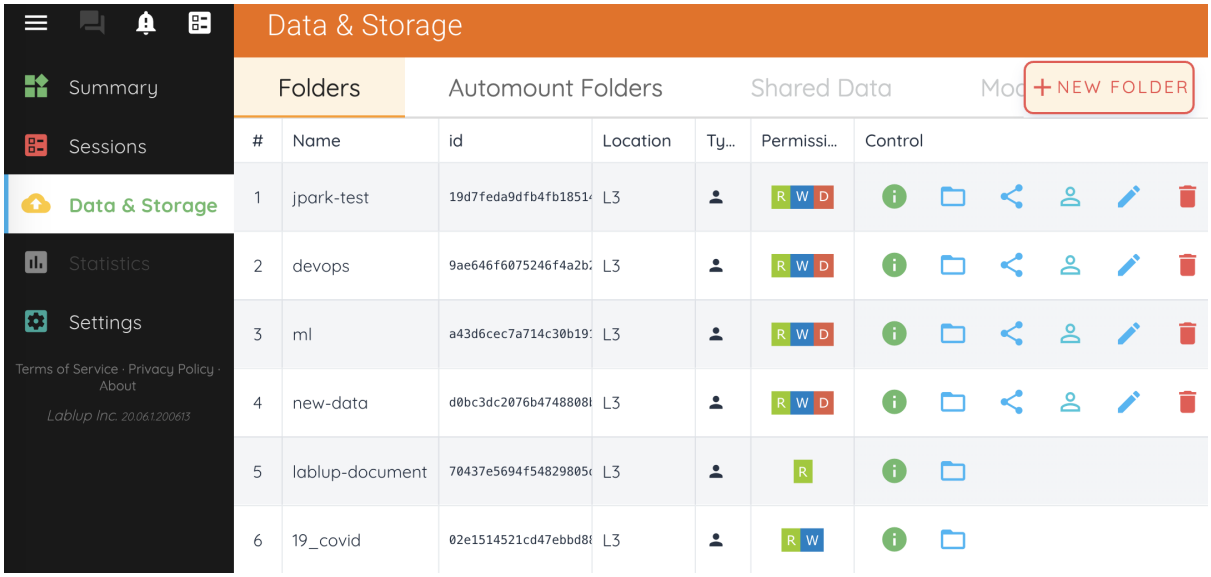
- `Ctrl-B c`: create a new tmux shell
- `Ctrl-B w`: query current tmux shells and move around among them
- `exit` or `Ctrl-B x`: terminate the current shell

Combining the above commands allows you to perform various tasks simultaneously on multiple shells.

CHAPTER 7

Handing Folders

Backend.AI supports dedicated storage to preserve user's files. Since the files and directories of a compute session are deleted upon session termination, it is recommended to save them in a storage folder. List of storage folders can be found by selecting the Storage on the left sidebar.

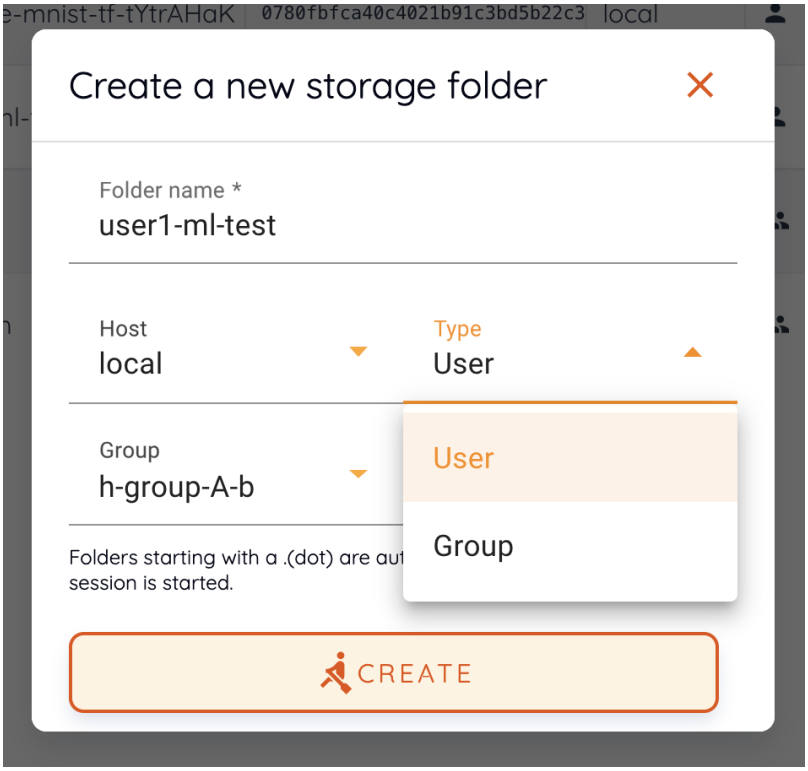


There are two types of storage folders. User type folders can be created by normal users,

and you can check that ther is one user icon in the Type column. On the other hand, Group folders can be recognized by an icon with multiple users in the column. Group folders are created by domain admins, and normal users can only see group folders which are created in a group where user belongs.

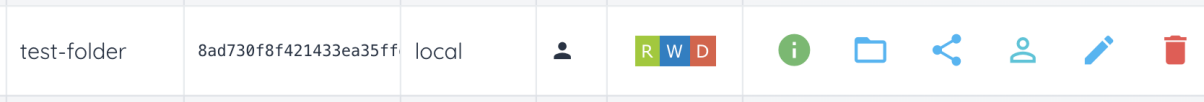
7.1 Create storage folder

You can create a storage folder with the desired name by clicking the NEW FOLDER button. Enter the name of the folder to be created in Folder name, and select one of User / Group for Type. (Depending on the server settings, only one of User or Group may be selectable.) When creating a group folder, the Group field must be set. The group folder will be bind to the group specified in the Group field, and only users belonging to the group can mount and use the group folder. After setting the values as desired, you can create a folder by clicking the CREATE button.

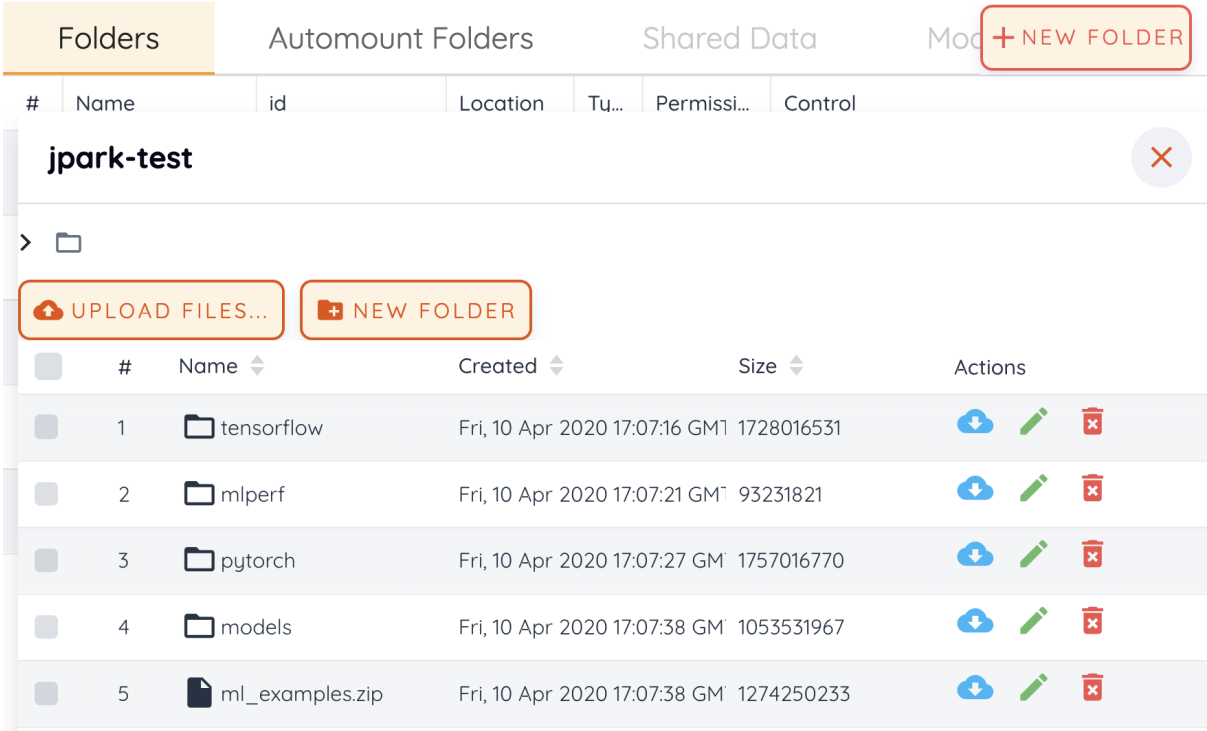


7.2 Explore folder

You can click the folder icon in the Control column to bring up a file explorer where you can view the contents of that folder.



You can see that directories and files inside the folder will be listed, if exists. Click a directory name in the Name column to move to the directory. You can click the download button or delete button in the Actions column to download it or delete it entirely from the directory. You can rename a file/directory as well. For more detailed file operations, you can mount this folder when creating a compute session, and then use a service like Terminal or Jupyter Notebook to do it.



You can create a new directory on the current path with the NEW FOLDER button (in the file explorer), or upload a local file with the UPLOAD FILES button. All of these file operations can also be performed using the above-described method of mounting folders into a compute session.

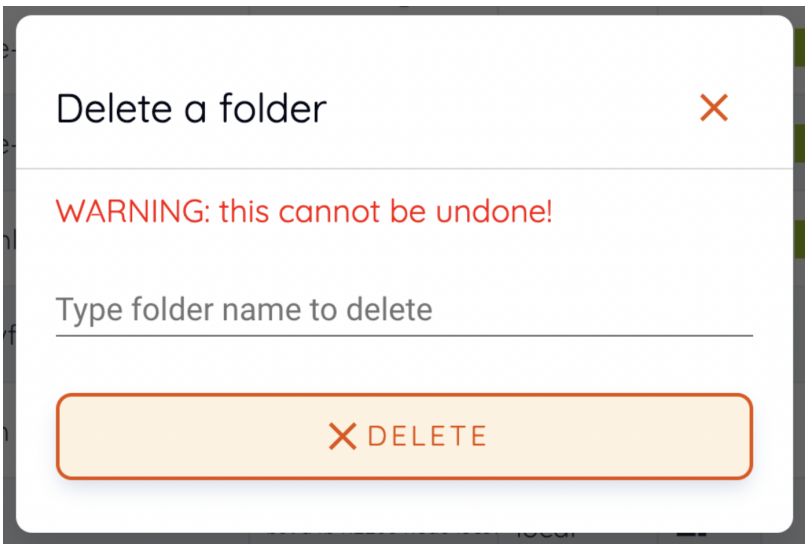
To close file explorer, click the X button in the upper right.

7.3 Rename folder

If you have permission to rename the storage folder, you can rename it by clicking the edit icon in the Control column. When you click the icon button, a rename dialog will appears. Write new folder name and then click RENAME button.

7.4 Delete folder

If you have permission to delete the storage folder, you can delete it by clicking the trash can icon in the Control column. When you click the Delete button, a confirm dialog appears. To prevent accidental deletion, you have to enter the name of the folder to be deleted, explicitly.

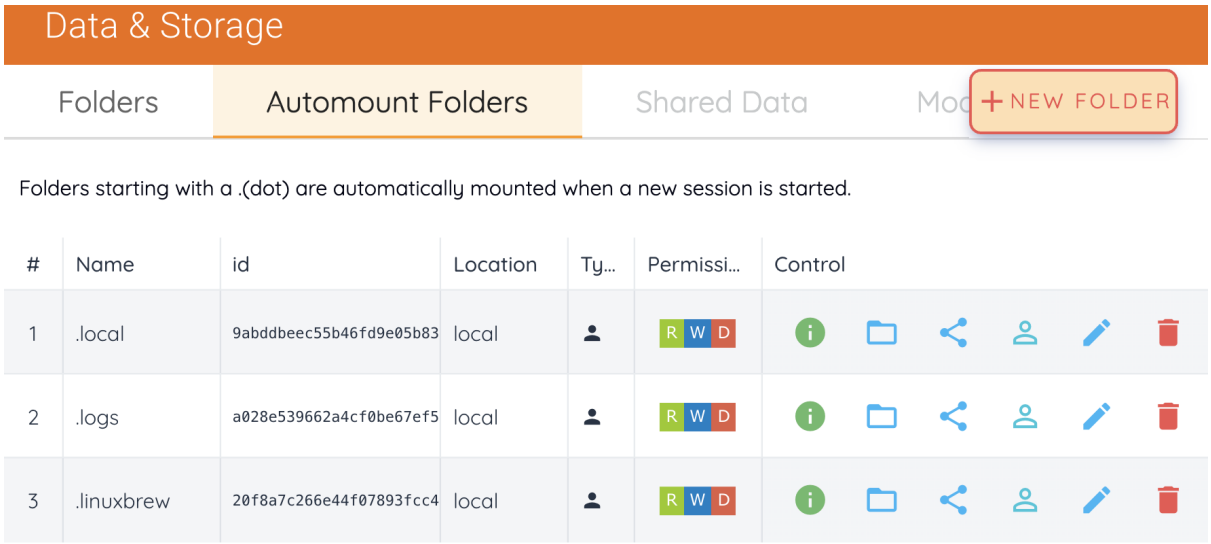


The folders created here can be mounted when creating a compute session. Folders are mounted under the user's default working directory, `/home/work/`, and the files stored in this directory will not be deleted when the compute session is terminated. (However, if you delete the folder itself, it will be gone).

7.5 Automount folder

Storage & Folders page has an Automount Folders tab. Click this tab to see a list of folders whose names prefixed with a dot (.). When you create a folder, if you specify a name that starts with a dot (.), it is added to the Automount Folders tab, not the Folders tab. Automount Folders are special folders that are automatically mounted in your home directory even if you do not mount them manually when creating a compute session. By using this feature, creating and using Storage folders such as `.local`, `.linuxbrew`, `.pyenv`, and etc., you can configure a certain user packages or environments that do not change with different kinds of compute session.

For more detailed information on the usage of Automount folders, refer to *examples of using automount folders*.

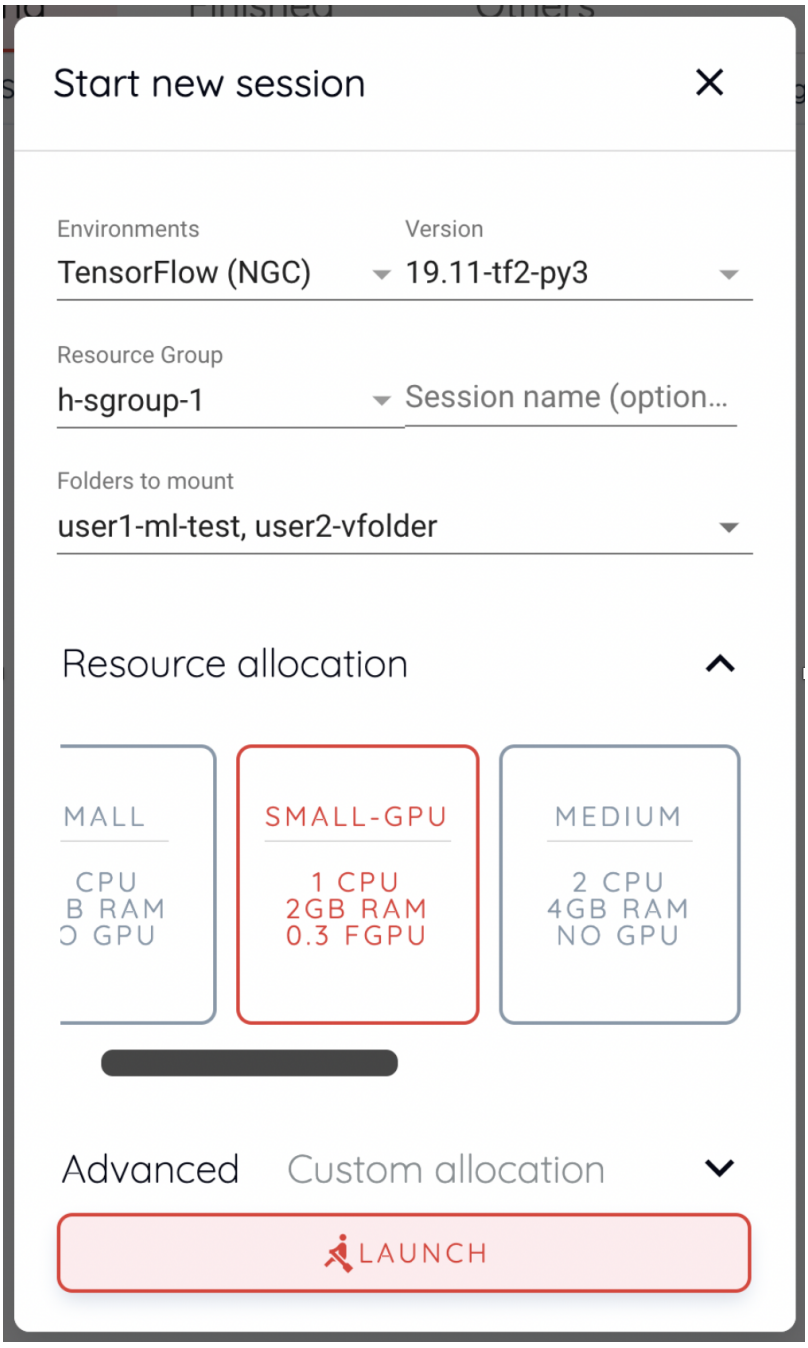


CHAPTER 8

Create a Compute Session with Mounted Folders

When you start a compute session, the user has access to the `/home/work/` directory, and the normal directores and files created under `/home/work/` will disappear when the compute session is destroyed. This is because compute sessions are dynamically created and deleted based on the container. To preserve the data inside the container independent of the life cycle of the container, a separate host folder must be mounted into the container, and then files must be created within the mounted folder. Backend.AI provides a function to mount storage folders when creating a compute session.

Let's go to the Sessions page and click the START button to create a new compute session. In the session create dialog, click Folders to mount to see a list of storage folders that a user can mount. Among them, you can add them by clicking the folder you want to mount. You can also mount multiple folders simultaneously by clicking multiple items. In this example, we will mount two folders, `user1-ml-test` and `user2-vfolder`, and then create a compute session.

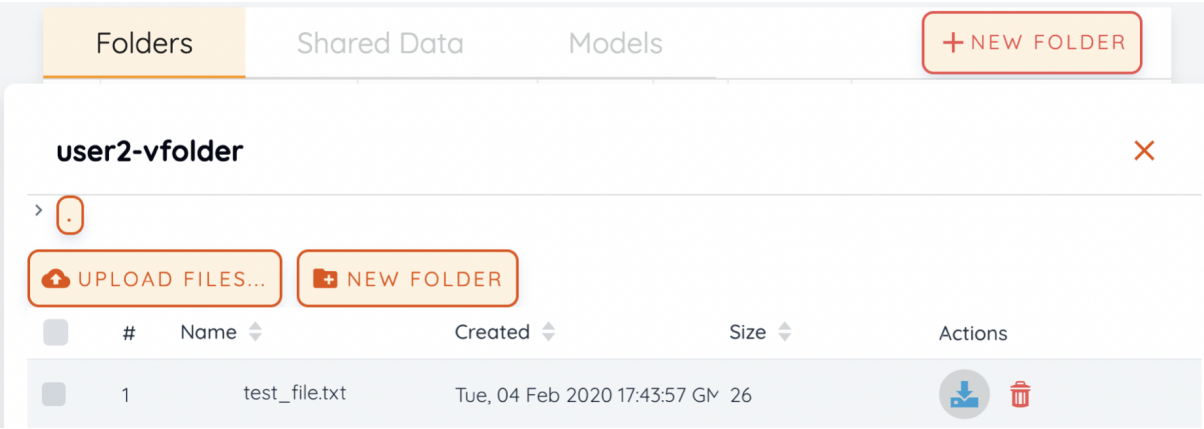


Now, open the terminal by clicking the terminal icon in the created session. If you run `ls` command in the terminal, you can see that the `user1-ml-test` and `user2-vfolder` folders are mounted under the home directory. Let's create a `test_file` under `user2-vfolder` to see if the file can be preserved after the compute session is destroyed. The contents of this file will filled with "file inside user2-vfolder".

```
work@7d51e0e00bbf:~$ ls
id_container  user1-ml-test  user2-vfolder
work@7d51e0e00bbf:~$ ls user2-vfolder/
basic_classification.ipynb  tc-perft01.txt
work@7d51e0e00bbf:~$ echo "file inside user2-vfolder" > user2-vfolder/test_file
work@7d51e0e00bbf:~$ ls user2-vfolder/
basic_classification.ipynb  tc-perft01.txt  test_file
work@7d51e0e00bbf:~$ cat user2-vfolder/test_file
file inside user2-vfolder
work@7d51e0e00bbf:~$
```

If you run `ls` command against `user2-vfolder`, you can see that the file was created successfully. Also note the contents of the file with the `cat` command.

Now delete the compute session and go to the Storage page. Locate the `user2-vfolder folder`, open a file explorer and check that the file `test_file` exists. Click the file download button in Actions to download the file to the local machine and open it in an editor to confirm that the contents of the file are "file inside user2-vfolder".



Like this, when creating a compute session, you can mount storage folders and perform any file operations on those mounted folders to save data even after the compute session termination.

8.1 Configuring a compute session environment using an automount folder

Sometimes you need a new program or library that is not pre-installed in a compute session. In that case, you can install packages and configure a certain environment

regardless of the type of compute session by using the Storage folder, which persists independent of session lifecycle, and the *automount folder*.

Install Python packages via pip

Creating a folder named `.local` allows a user to install Python user packages in that folder. This is because installing a package with the `-user` option appended to `pip` installs the package in the `.local` folder under the user's home folder (note that automount folder is mounted under user's home folder). So, if you want to install and keep the Python package `tqdm` regardless of the type of computing environment, you can issue the following command in your terminal:

```
pip install --user tqdm
```

After that, when a new compute session is created, the `.local` folder where the `tqdm` package is installed is automatically mounted, so you can use the `tqdm` package without reinstalling.

Warning: If you spawn multiple sessions that uses multiple Python versions, there may be compatibility issues with the packages. This can be circumvented by branching `PYTHONPATH` environment variable via the `.bashrc`. This is because the user's `pip` package is installed in the path specified in the `PYTHONPATH`.

Install packages via Homebrew

Package managers like Ubuntu's `apt` or CentOS's `yum` usually require `sudo` permissions. For security, `sudo` and `root` accesses are blocked by default in Backend.AI's compute session (it may allowed depending on the configuration), so we recommend to use Homebrew on Linux which does not require `sudo`. Homebrew can be installed and used in the following ways:

- Create `.linuxbrew` folder in Data & Storage page

- Create a compute session (`.linuxbrew` folder is automatically mounted)
- Install Homebrew on Linux

```
sh -c "$(curl -fsSL https://raw.githubusercontent.com/Linuxbrew/install/master/
↪install.sh)"
export PATH=/home/work/.linuxbrew/bin:$PATH
brew
```

- Install package

```
brew install fortune
fortune
```

You can manage various settings using the automount folder in the same way as above.

More details can be found on the Backend.AI wiki.

CHAPTER 9

Connect to a Compute Session with SSH/SFTP

Backend.AI supports SSH/SFTP connection to the created compute sessions (containers). In this section, we will lear how.

Note: SSH/SFTP connection is supported only on desktop apps, and not yet supported on web-based console service.

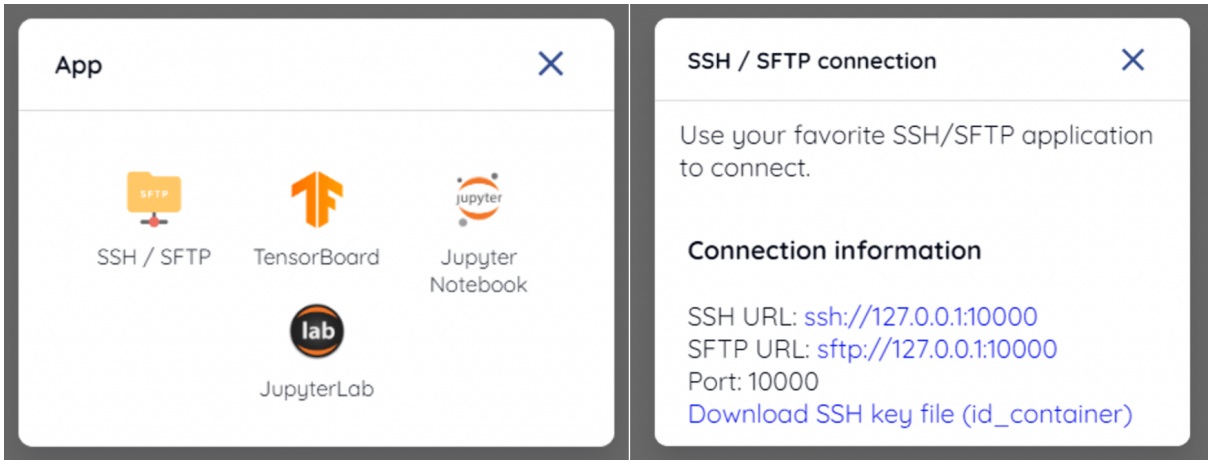
9.1 SSH/SFTP connection to a compute session (Linux/Mac)

First, create a compute session, then click the app icon (first button) in Control, followed by SSH / SFTP icon. Then, a daemon that allows SSH/SFTP access from inside the container will be initiated, and the Console app interacts with the daemon through

a local proxy service.

> **Warning:** You cannot establish a SSH/SFTP connection to the session until you click the SSH/SFTP icon. When you close the Console app and launch it again, the connection between the local proxy and the Console app is initialized, so the SSH/SFTP icon must be clicked again.

Next, a dialog containing SSH/SFTP connection information will be pop up. Remember the address (especially the assigned port) written in the SFTP URL and click the download link to save the `id_container` file on the local machine. This file is an automatically generated SSH private key. Instead of using the link, you can also download the `id_container` file located under `/home/work/` with your web terminal or Jupyter Notebook. The auto-generated SSH key may change when new session is created. In that case, it must be downloaded again.



To SSH connect to the compute session with the downloaded SSH private key, you can run the following command in the shell environment. You should write the path to the downloaded `id_container` file after `-i` option and the assigned port number after `-p` option. The user inside the compute session is usually set to `work`, but if your session uses other account, the `work` part in `work@localhost` should be changed to the actual session account. If you run the command correctly, you can see that SSH connection is made to the compute session and you are welcomed by the container's

shell environment.

```
$ ssh -o StrictHostKeyChecking=no \
>     -o UserKnownHostsFile=/dev/null \
>     -i ~/.ssh/id_container \
>     work@localhost -p 10022
Warning: Permanently added '[127.0.0.1]:9922' (RSA) to the list of known hosts.
f310e8dbce83:~$
```

Connting by SFTP would almost be the same. After running the SFTP client and setting public key-based connection method, then simply specify `id_container` as the SSH private key. Each FTP client may adopt different way, so refer to each FTP client manual for details.

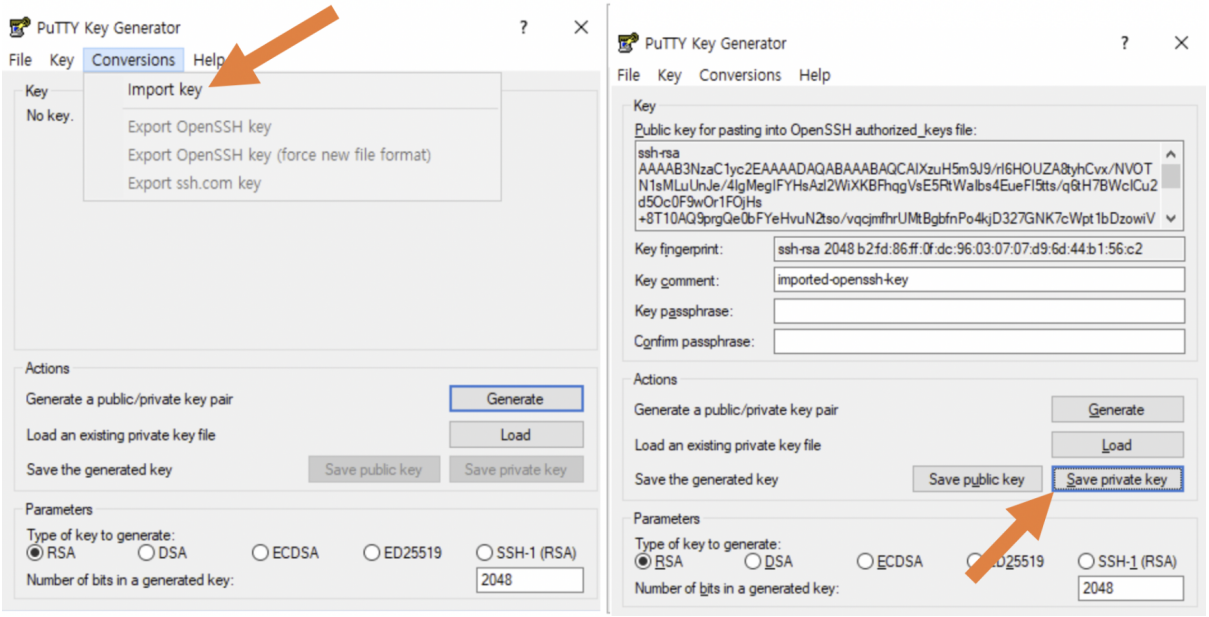
Note: The SSH/SFTP connection port number is randomly assigned each time a session is created. If you want to use a specific SSH/SFTP port number, you can input the port number in the "Preferred SSH Port" field in the user settings menu. To avoid possible collisions with other services within the compute session, it is recommended to specify a port number between 10000-65000. However, if SSH/SFTP connections are made by two or more compute sessions at the same time, the second SSH/SFTP connection cannot use the designated port (since the first SSH/SFTP connection is already taken it), so a random port number will be assigned.

Note: If you want to use your own SSH keypair instead of `id_container`, create a user-type folder named `.ssh`. If you create `authorized_keys` file in that folder and append it with the contents of your SSH public key, you can connect by SSH/SFTP through your own SSH private key without having to download the `id_container` after creating the compute session. There is.

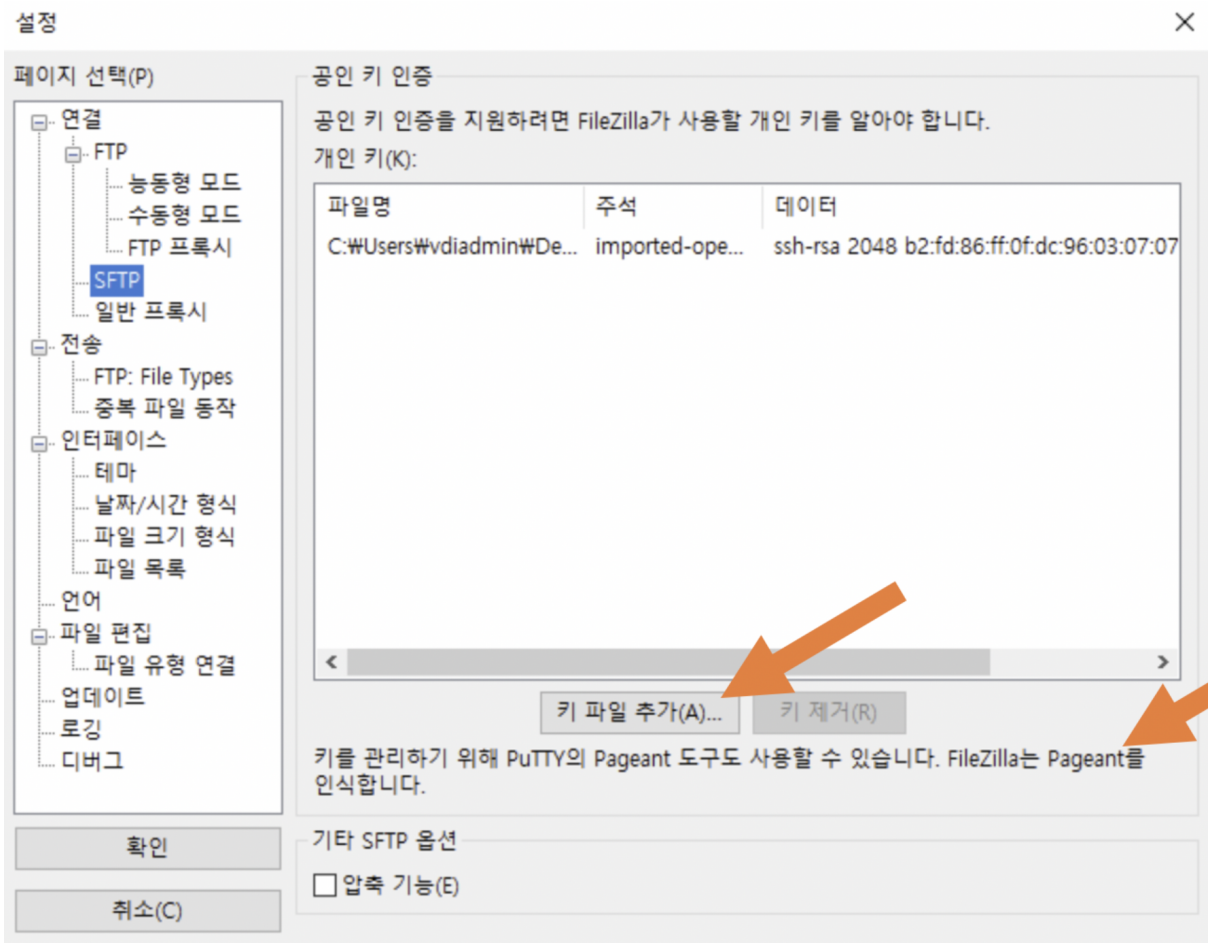
9.2 SSH/SFTP connection to a compute session (Windows / FileZilla)

Backend.AI Console app supports OpenSSH-based public key connection (RSA2048). To access with a client such as PuTTY on Windows, a private key must be converted into a ppk file through a program such as PuTTYgen. You can refer to the following link for the conversion method: https://wiki.filezilla-project.org/Howto. For easier explanation, this section will describe how to connect to SFTP through FileZilla client on Windows.

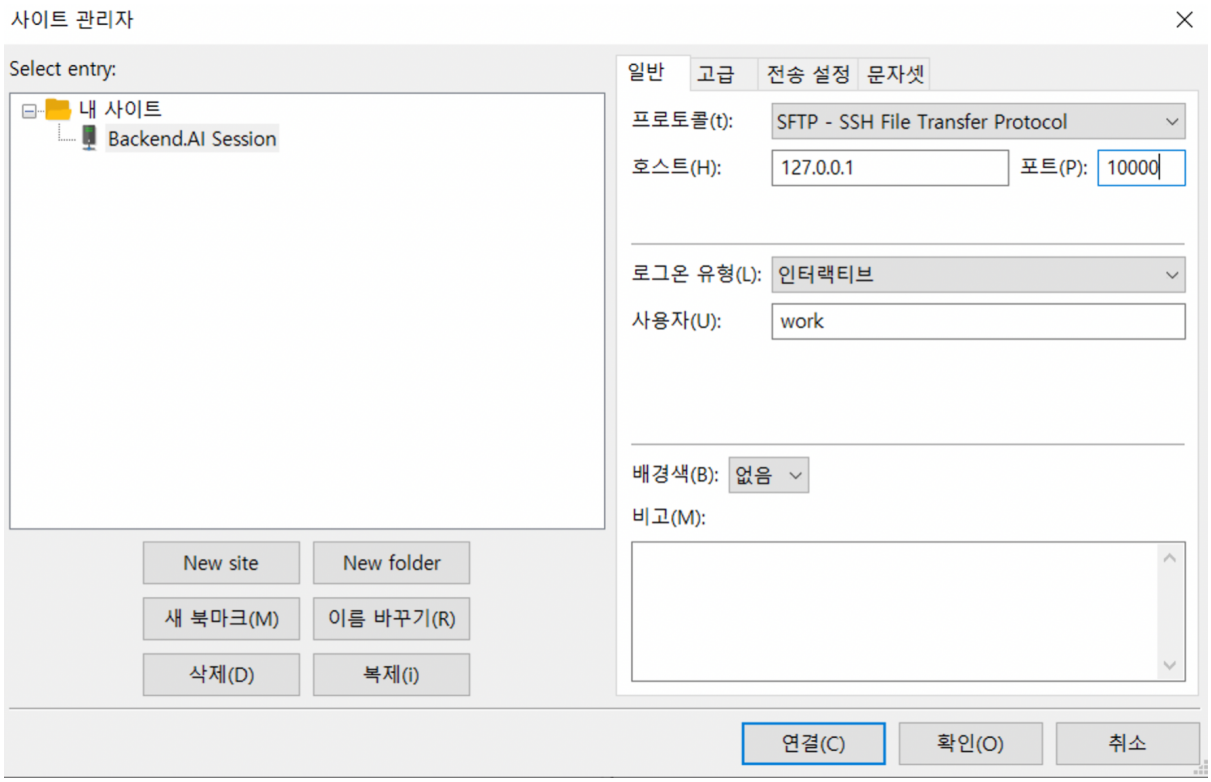
Refer to the connection method on Linux/Mac, create a compute session, check the connection port and download `id_container`. `id_container` is an OpenSSH-based key, so if you use a client that supports only Windows or ppk type keys, you must convert it. Here, we will convert through the PuTTYgen program installed with PuTTY. After running the PuTTYgen, click on the import key in the Conversions menu. Select the downloaded `id_container` file from the file open dialog. Click the Save private key button of PuTTYGen and save the file with the name `id_container.ppk`.



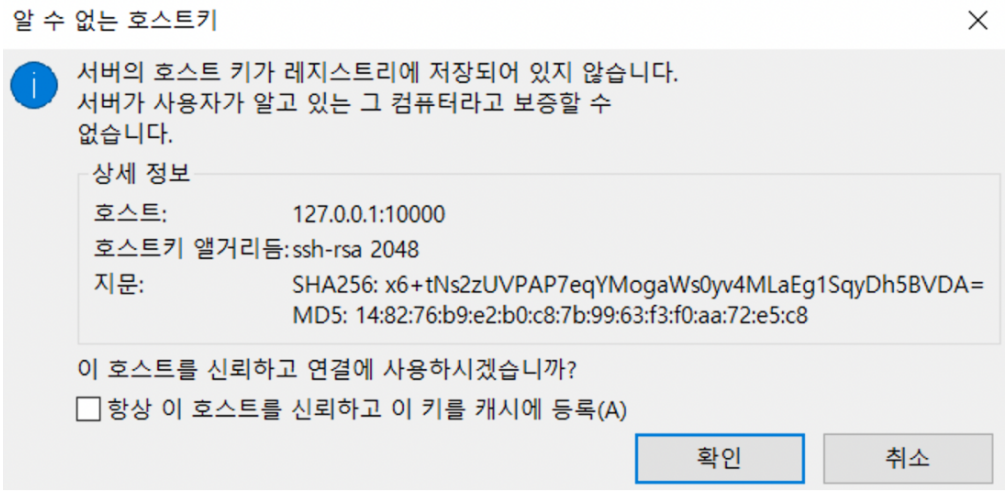
After launching the FileZilla client, go to the Settings-Connection-SFTP and register the key file `id_container.ppk` (`id_container` for clients supporting OpenSSH).



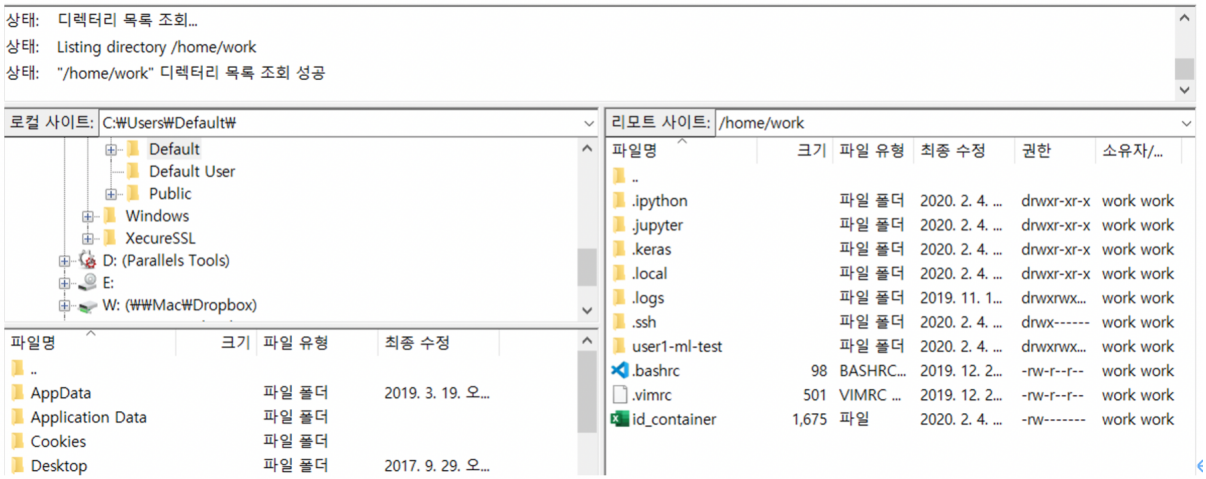
Open Site Manager, create a new site, and enter the connection information as follows.



When connecting to a container for the first time, the following confirmation popup may appear. Click the OK button to save the host key.



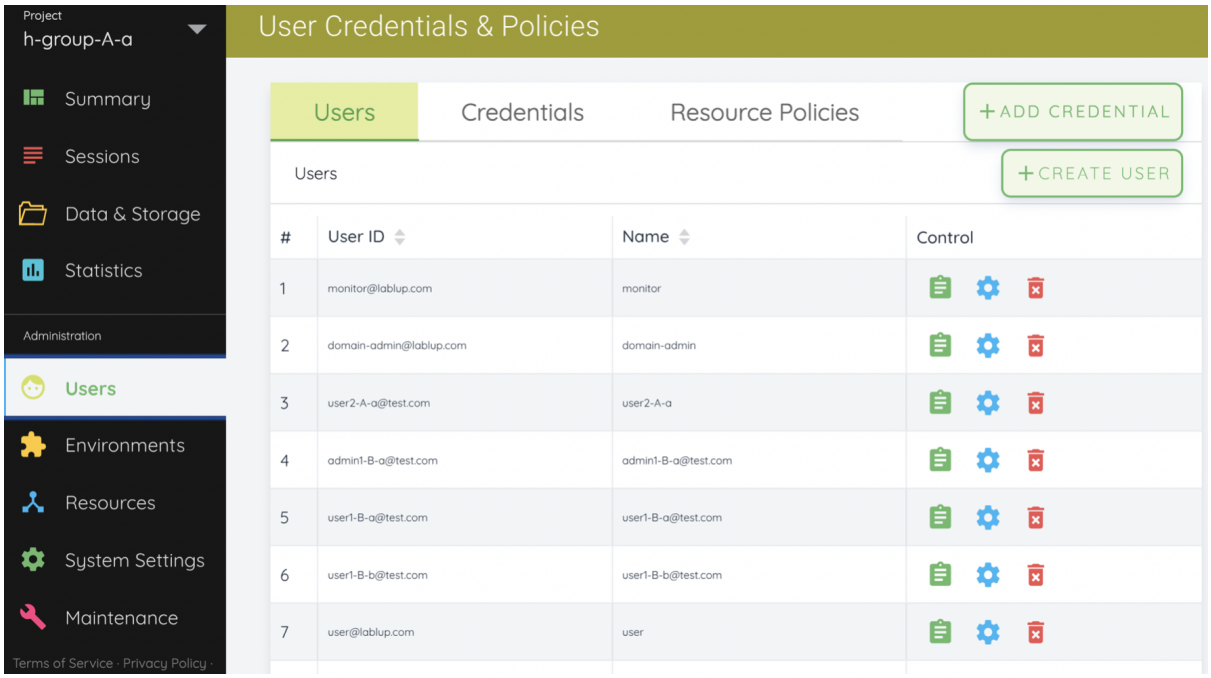
After a while, you can see that the connection is established as follows. You can now transfer large files to `/home/work/` or other mounted storage folder with this SFTP connection.



CHAPTER 10

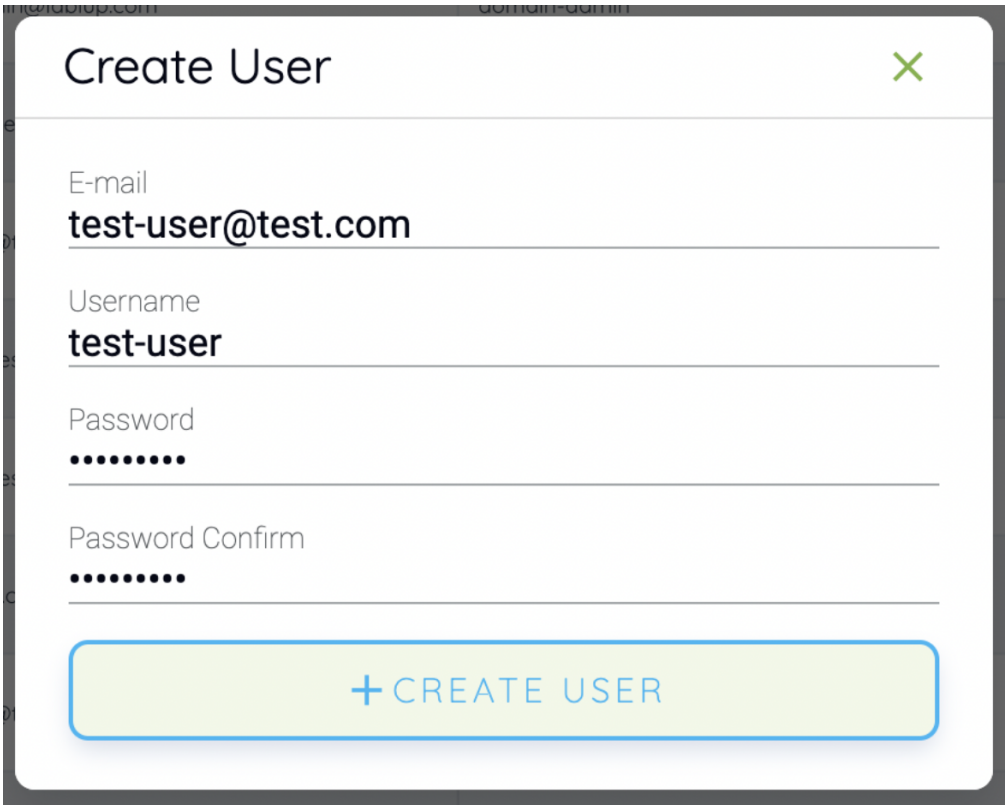
Admin Menus

When you log in with an admin account, you will see an additional Administration menu in the bottom left sidebar. User information registered in Backend.AI is listed in the Users tab. In the case of domain admin, users belong to the domain can be only be listed, while superadmin can see all user information. Only superadmin can create and deactivate a user.

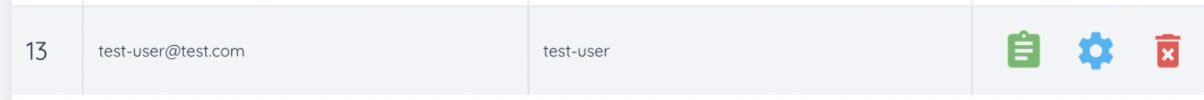


10.1 Create and update user

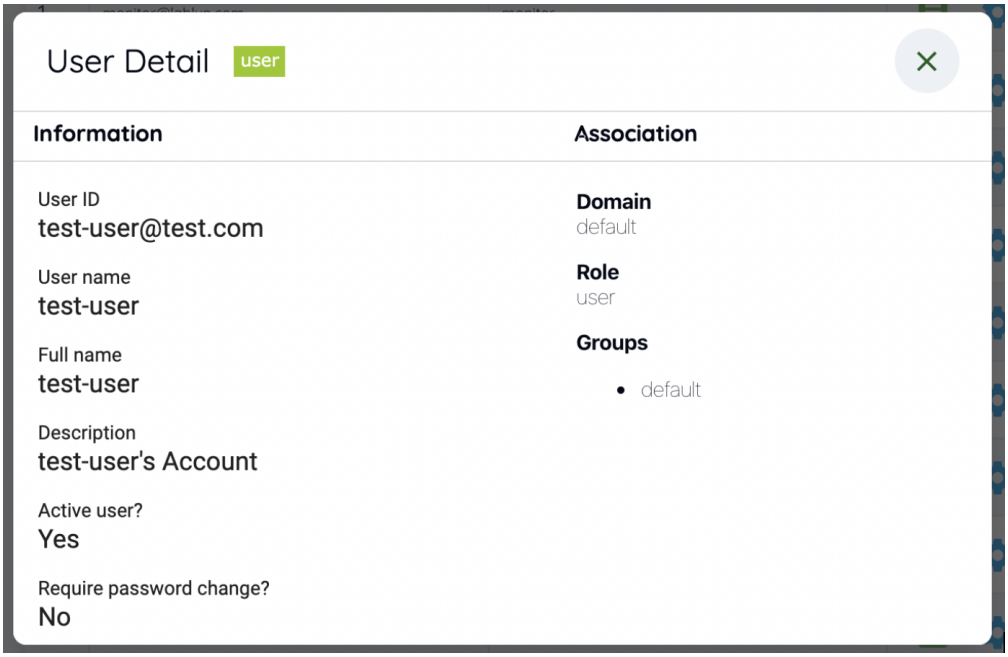
A user can be created by clicking the CREATE USER button. Note that the password must be longer or equal to 8 characters and at least 1 alphabet/special character/ number should be included.



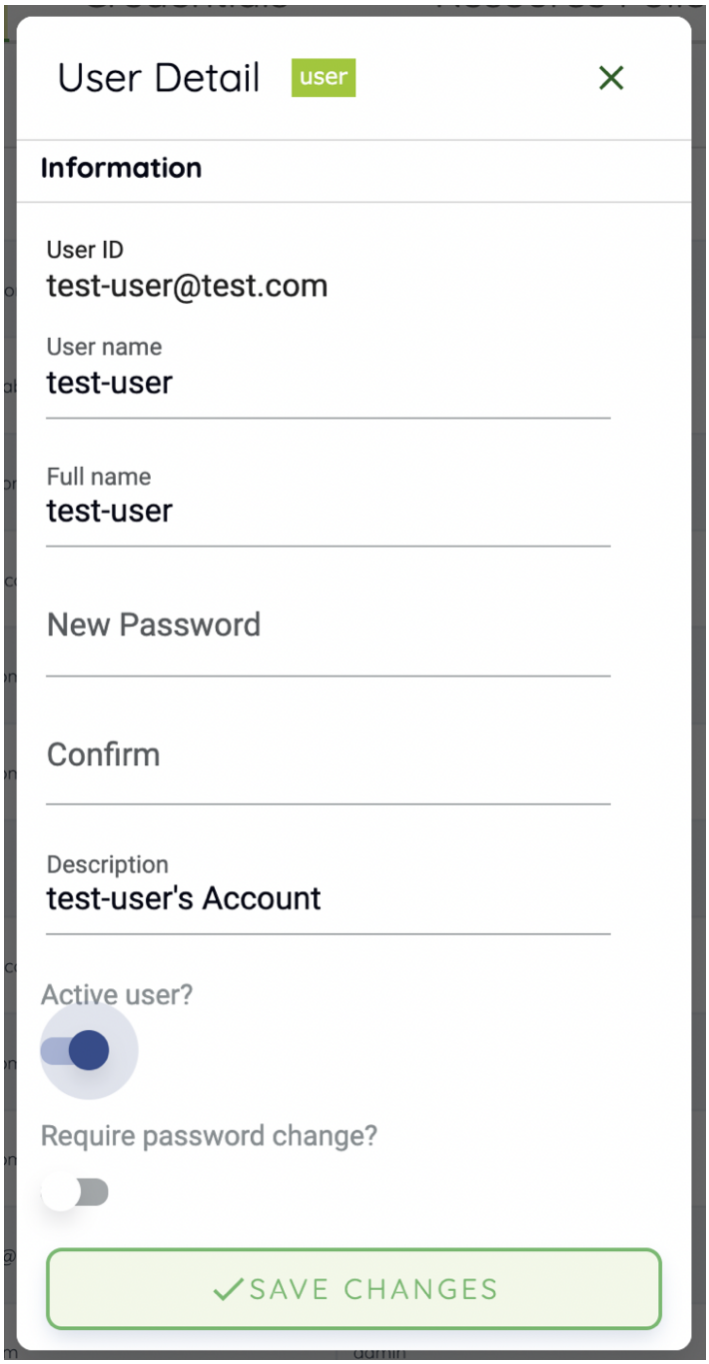
Check if the user is created.



Click the green button in the Controls column for more detailed user information. You can also check the domain and group information where the user belongs.



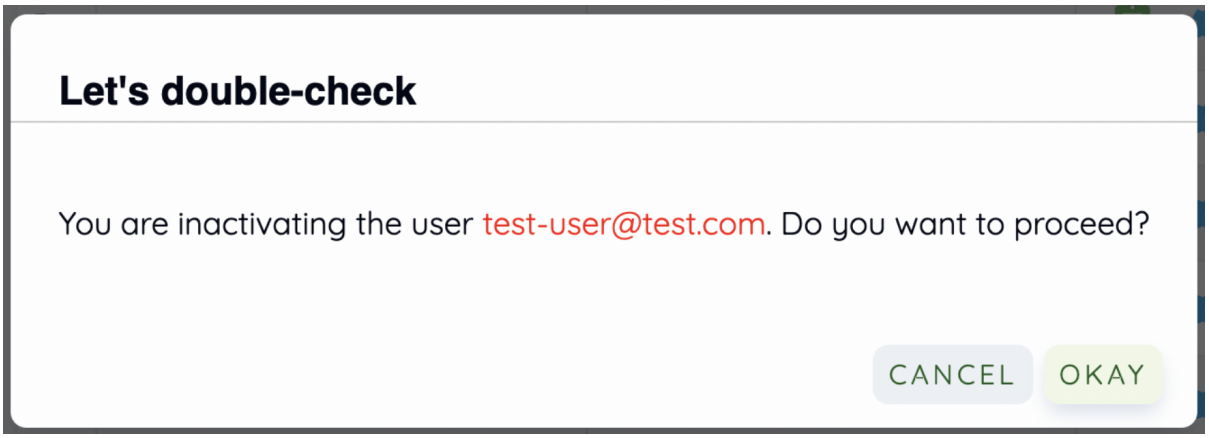
Click the gear icon in the Controls column to update information of a user who already exist. User's name, password, activation state, and etc. can be changed.



10.2 Deactivate user account

Deleting user accounts is not allowed even for superadmins, to track usage statistics per user, metric retention, and accidental account loss. Instead, admins can deactivate user

accounts to keep users from logging in. Click the trash icon in the Controls column. A dialog asking confirmation appears, and you can deactivate the user by clicking the OKAY button.



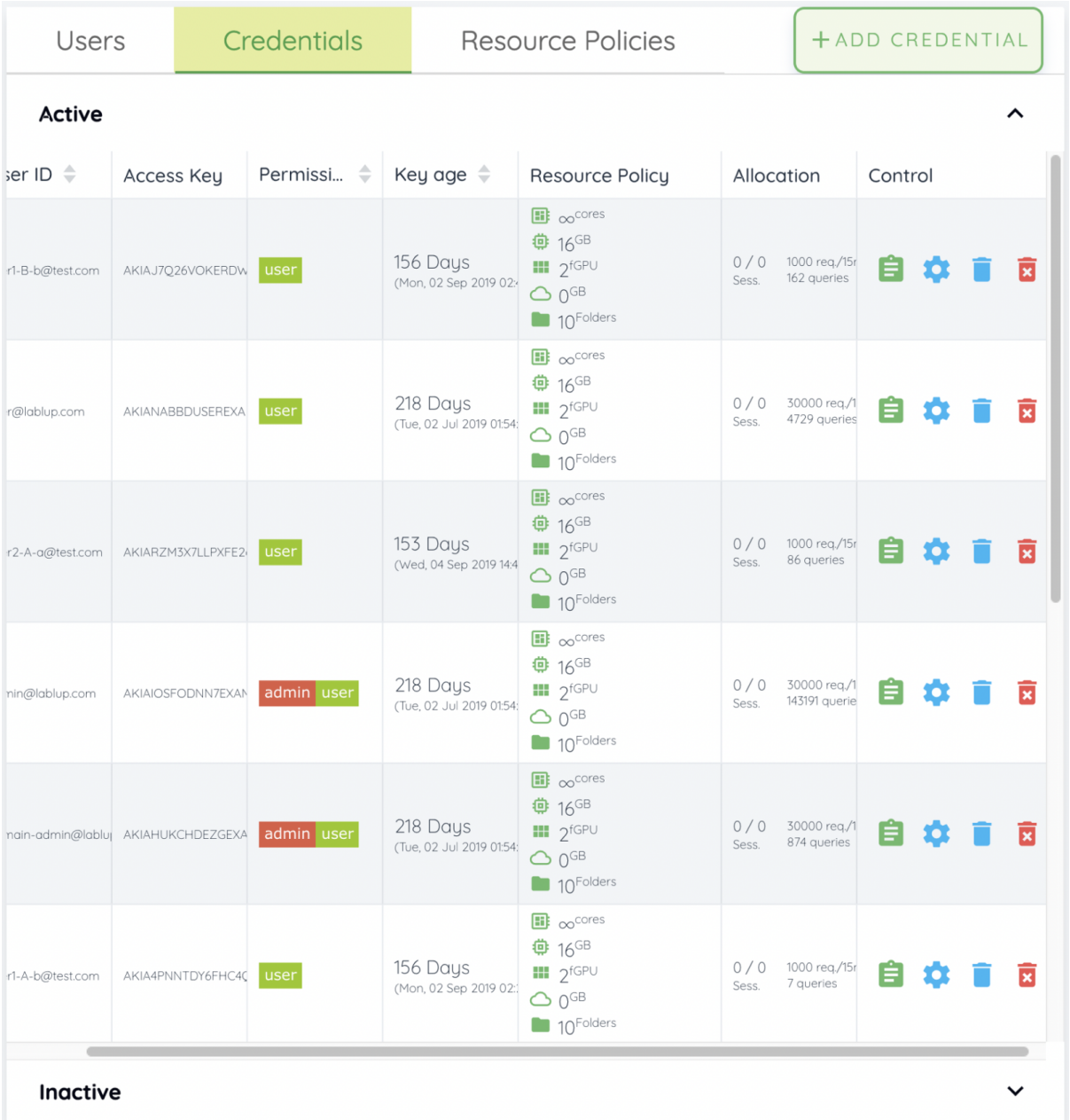
Deactivated users are not listed in Users tab.

10.3 Manage User's Keypairs

Each user account usually have one or more keypairs. A keypair is used for API authentication to the Backend.AI server, after user logs in. Login requires authentication via user email and password, but every request the user sends to the server is authenticated based on the keypair.

A user can have multiple key pairs, but to reduce users the burden of managing keypair, we are currently using only one of user's keypairs to send requests. Also, when you create a new user, a keypair is automatically created, so you do not need to create and assign a keypair manually in most cases.

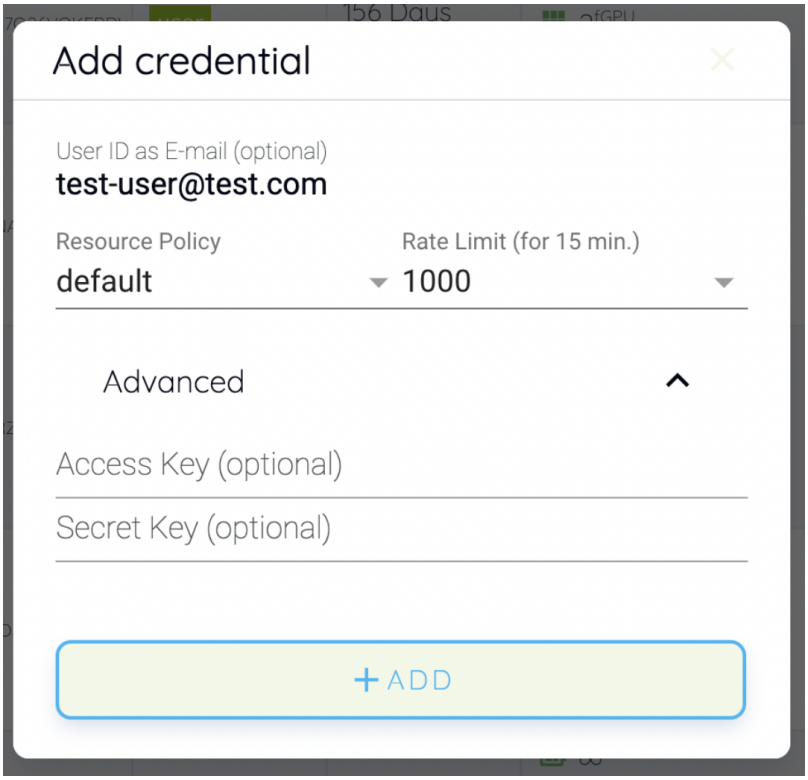
Keypairs can be listed on the Credentials tab of in the Users page. Active keypairs are shown immediately, and to see the inactive keypairs, click the Inactive panel at the bottom.



Like in Users tab, you can use the buttons in the Controls column to view or update keypair details. Click the blue trash button to disable that key pair, or click the red trash button to completely delete the keypair. However, if you have created a compute session using a key pair, you cannot delete it. If you accidentally deleted a keypair, you can re-create keypair for the user by clicking the ADD CREDENTIAL button at the upper right corner. If necessary, you can also explicitly enter the access key and secret key by clicking the Advanced panel.

The Rate Limit field is where you specify the maximum number of requests that can

be sent to the Backend.AI server in 15 minutes. For example, if set to 1000, and the keypair sends more than 1000 API requests in 15 minutes, and the server throws an error and does not accept the request. It is recommended to use the default value and increase it when the API request frequency goes up high according to the user's pattern.

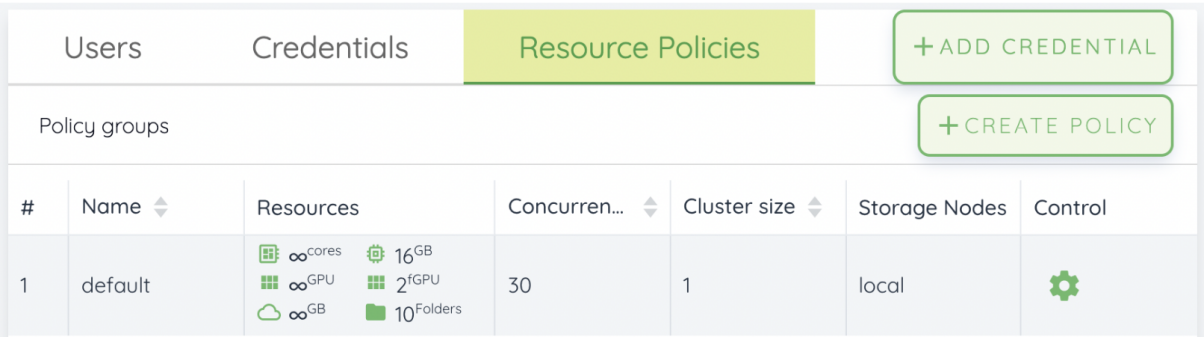


10.4 Manage Resource Policy

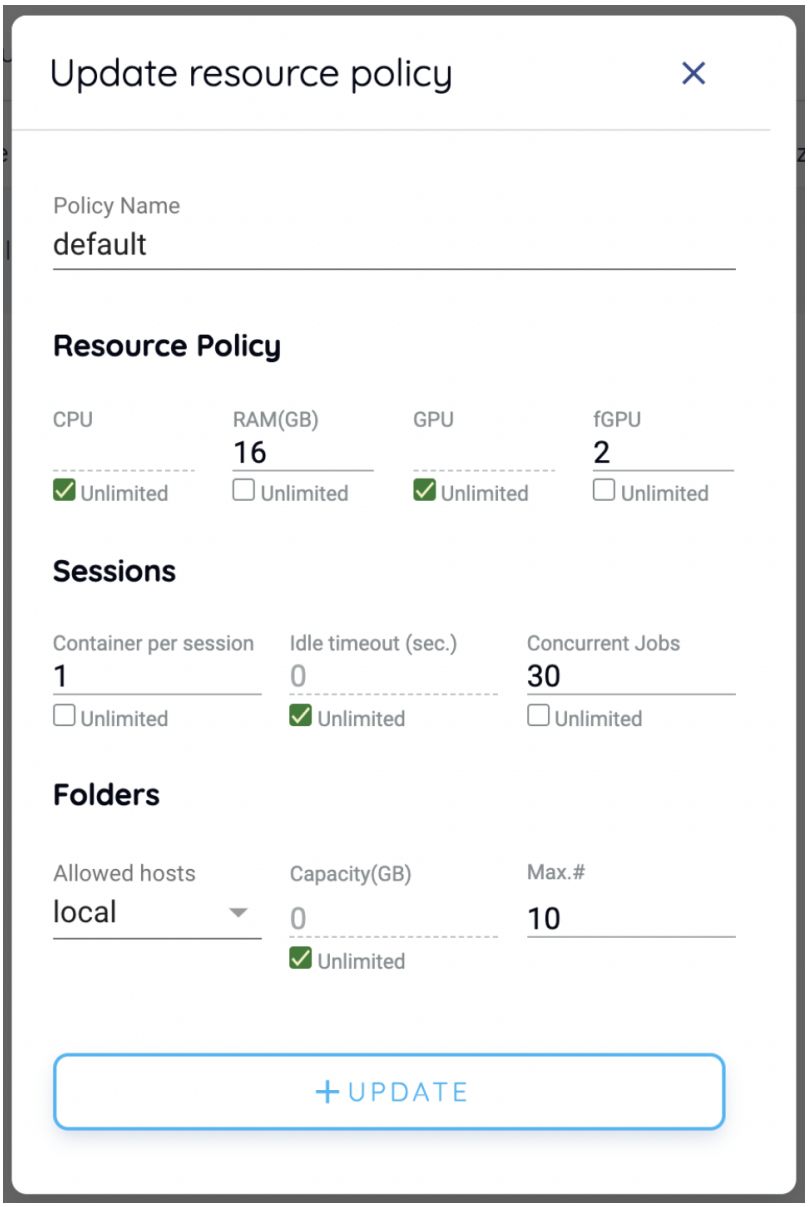
Resource policies can be listed and modified in Resource Policies tab on the Users page. Resource policies allow you to set maximum allowed resources and/or other compute session related settings per keypair basis. If necessary, multiple resource policies can be created, for example, user / research purposes, and apply them separately to each keypair.

Note: To set resource policies at the domain and group level, you have to use Manager Hub, a UI for superadmin only. In the Console UI, you can only set resource policies

based on keypairs. Manager Hub is part of the enterprise version.



In the example image above, there is one resource policy named `default`. You can change the resource policy by clicking the settings icon in the Control column. After changing the settings, click UPDATE to save.



The meaning of each field is as follows.

- CPU: The maximum number of CPU cores that a keypair can use. For example, if set to 4, keypairs bound to the resource policy cannot assign more than 4 cores to the container. Note that the number of cores is limited based on the sum of all containers created by the keypair. If one container allocates three cores, a new container can only allocate one core. In addition, if you check Unlimited, the keypair can allocate resources as much as the server permits. This also applies to other resource settings.
- RAM: Maximum memory.

- GPU: The maximum number of physical GPUs that can be allocated. Used only when the GPU provisioning mode of the Backend.AI server is set to "device".
- fGPU: The maximum number of virtualized GPUs that can be allocated. Used only when the GPU provisioning mode of the Backend.AI server is set to "shares". The unit of fGPU is independent of the number of physical GPU devices, and is determined by the streaming multiprocessor (SM) and GPU memory unit set by the server.
- Container per session: Maximum number of containers a session can have. This is a setting which will be used to bundle multiple containers and use them as a single session. The ability to bundle multiple containers is under development and is currently not used.
- Idle timeout: If a running session is not used for the time specified in the idle timeout, the session is automatically garbage collected (destroyed). You can set the time interval here. For example, if set to 600, sessions with no usage for 10 minutes are automatically destroyed. If set to 0 or checked Unlimited, garbage collection is not performed for the session created by the keypair.
- Concurrent Jobs: The maximum number of sessions a user can create concurrently. If set to 5, a keypair using that policy cannot create more than 5 compute sessions.
- Allowed hosts: Used to control the accessible storage and/or NFS hosts from a session when multiple storage/NFS hosts are available. Even if a NFS is mounted and can be used from Backend.AI, a user will not be able to use that host unless specified here. However, the NFS host may be accessible in case the host is configured to be accessible in the domain and/or group level. Domain / group level settings are possible in Manager Hub.
- Capacity: This is where you set the maximum available storage size. The disk size limit is only available under certain circumstances and is currently not supported.

This feature is under active development, and will be supported in the near future.

- Max. #: The maximum number of storage folders that can be created.

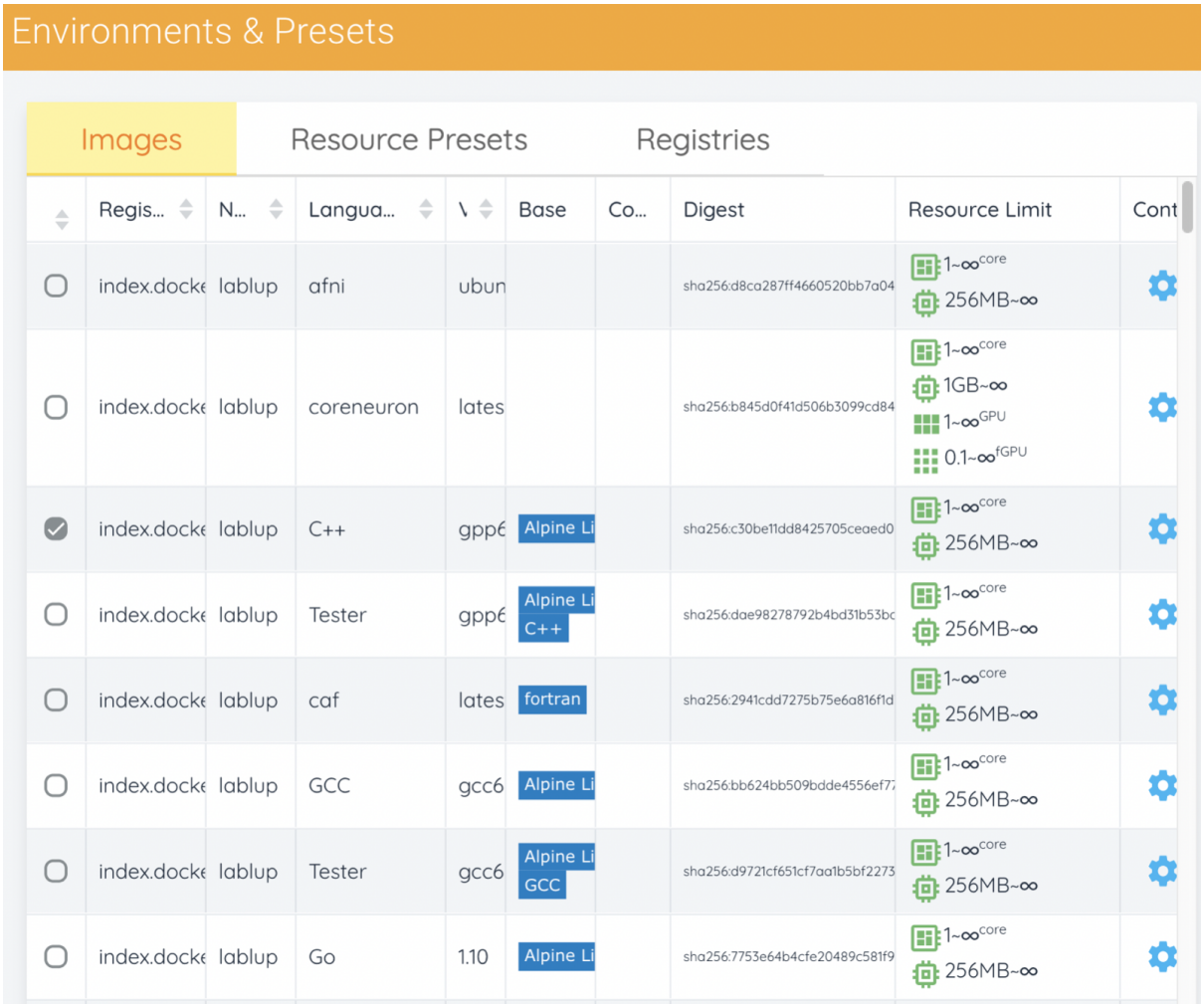
You can create a new resource policy by clicking the CREATE POLICY button. Each setting value is the same as described above.

To create a resource policy and associate it with a keypair, go to the Credentials tab of the Users page, click the settings button located in the Controls column of the desired keypair, and click the Select Policy field to choose it.

10.5 Manage Images

Admins can manage images, which are used in creating a compute session, in the Images tab of the Environments page. In the tab, meta information of all images currently in the Backend.AI server is displayed. You can check information such as registry, namespace, image name, image's based distros, digest, and minimum resources required for each image. For images downloaded to one or more agent nodes, a check mark is displayed on the left. An unchecked image means that it is not installed on any agent.

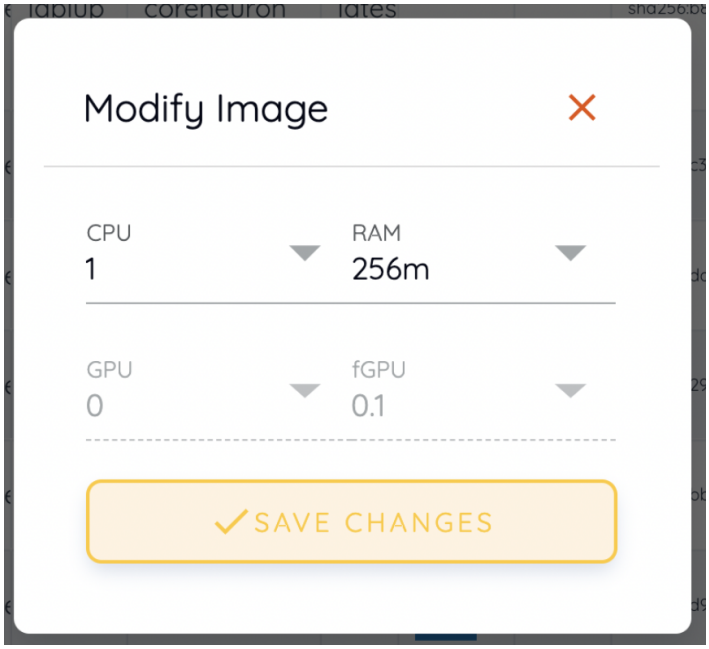
Note: The feature to install images by selecting specific agents is currently under development.



You can change the minimum resource requirements for each image by clicking the settings icon in the Controls column. Each image has hardware and resource requirements for minimal operation. (For example, for GPU-only images, there must be a minimum allocated GPU.) The default value for the minimum resource amount is provided as embedded in the image's metadata. If an attempt is made to create a compute session with a resource that is less than the amount of resources specified in each image, the request is automatically adjusted to the minimum resource requirements for the image and then generated, not canceled.

Warning: Don't change the minimum resource requirements to an amount less than the predefined value! The minimum resource requirements included in the image metadata are values that have been tested and determined. If you are not

really sure about the minimum amount of resources you want to change, leave it in the default.

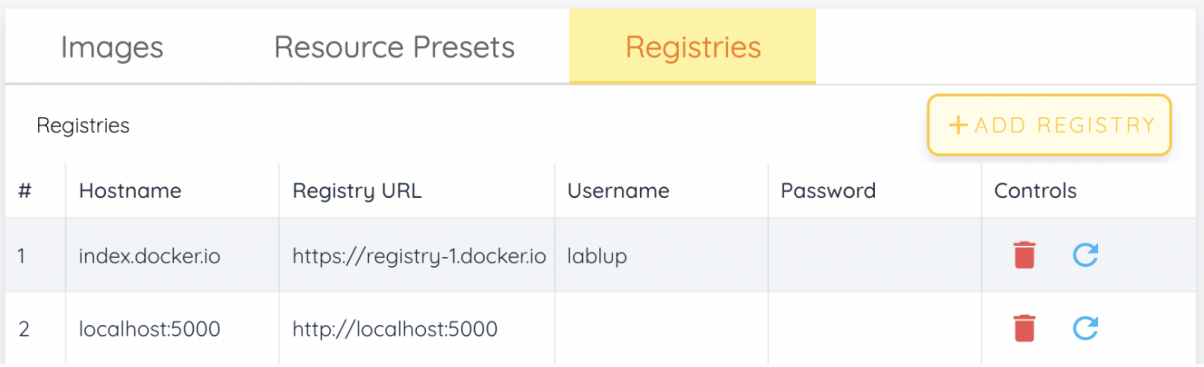


10.6 Manager docker registry

You can click on the Registries tab in Environments page to see the information of the docker registry that are currently connected to. `index.docker.io` is registered by default, and it is a registry provided by Docker.

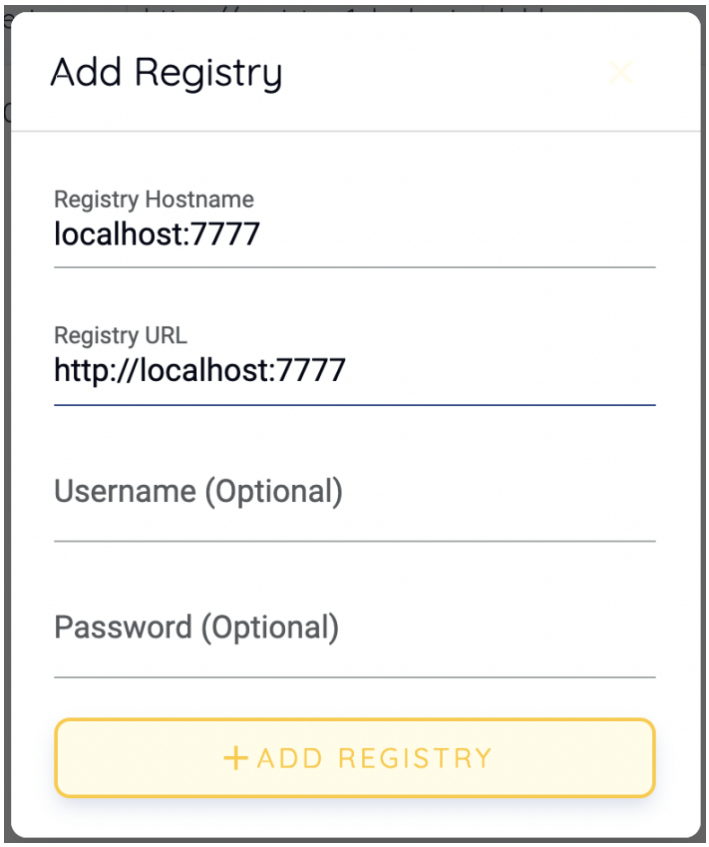
Note: In the offline environment, the default Docker registry is not accessible, so click the trash icon on the right to delete it.

Click the refresh icon in Controls to update image metadata for Backend.AI from the connected registry.AI. Image information which does not have labels for Backend.AI among the images stored in the registry is not updated.



You can add your own private docker registry by clicking the ADD REGISTRY button. Note that Registry Hostname and Registry URL address must be set identically, and in the case of Registry URL, a scheme such as `http://` or `https://` must be explicitly attached. Also, images stored in the registry must have a name prefixed with Registry Hostname. Username and Password are optional and can be filled in if you set separate authentication settings in the registry.

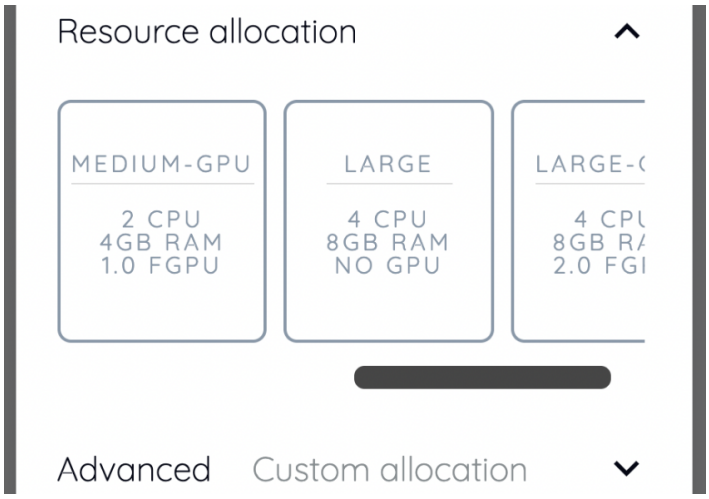
Note: In the case of `index.docker.io`, the Hostname and Registry URL are different, because Docker internally handles an exception for the default registry. If it is not the default registry, you must match the Hostname and Registry URL to properly connect.



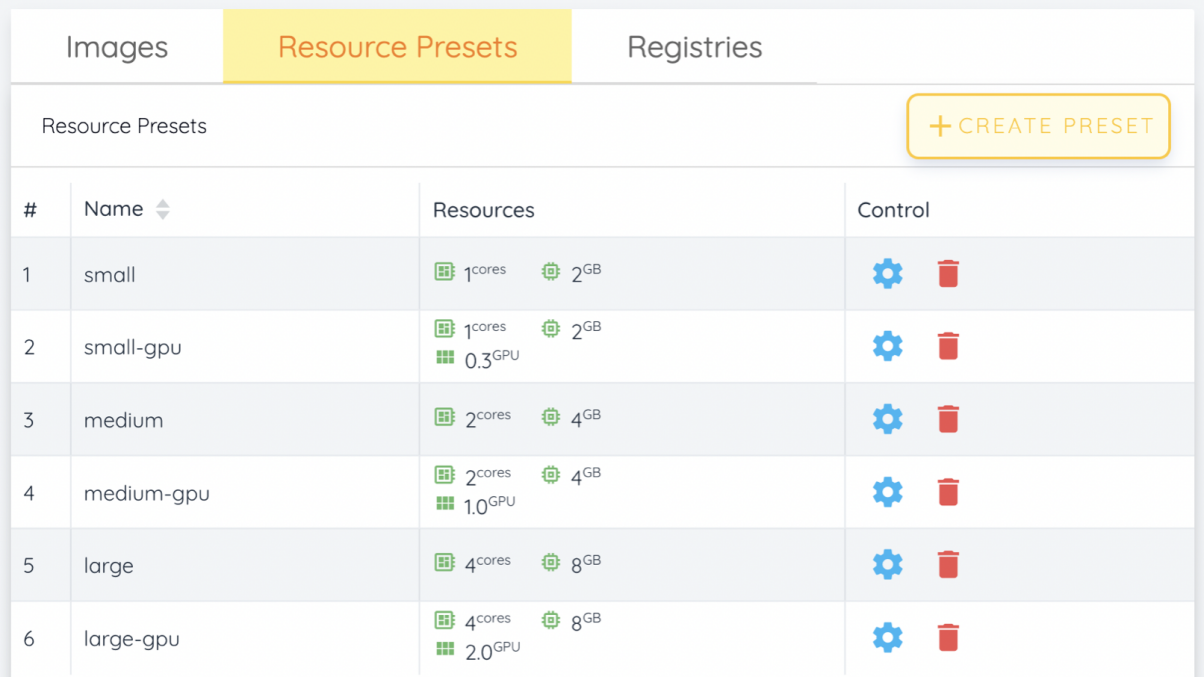
Even if you created a registry and update meta information, users cannot use the images in the registry, immediately. Just as you had to register allowed hosts to use the storage host, you must register the registry in the allowed docker registries field at the domain or group level, after registering the registry, so that users in the domain or group can access the registry image. Allowed docker registries can be registered using the Manager Hub with domain and group management. The ability to set allowed docker registries in the keypair's resource policy is not yet provided.

10.7 Manage resource preset

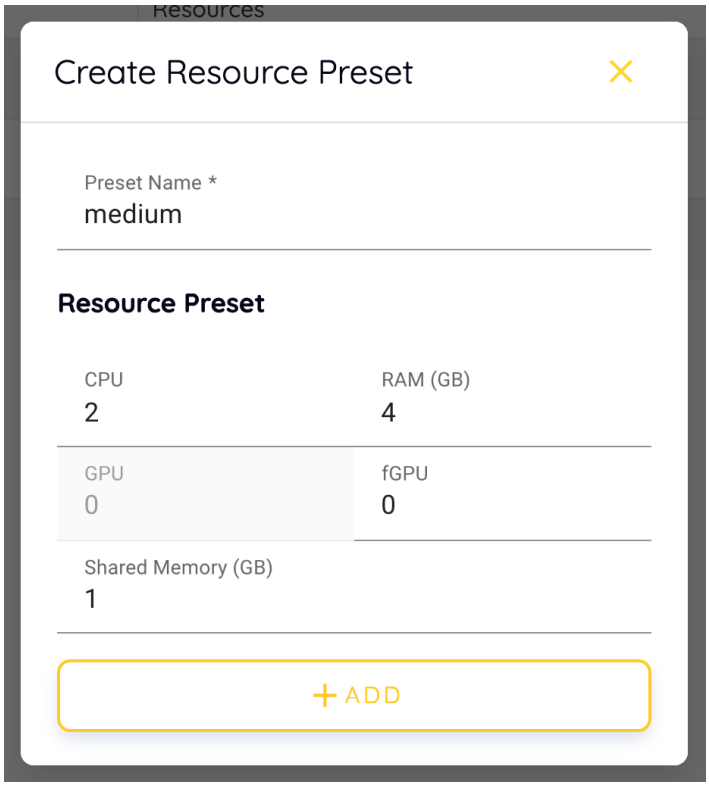
The following predefined resource presets are displayed in the Resource allocation panel when creating a compute session. Superadmin can manage these resource presets.



Go to the Resource Presets tab on the Environment page. You can check the list of currently defined resource presets.

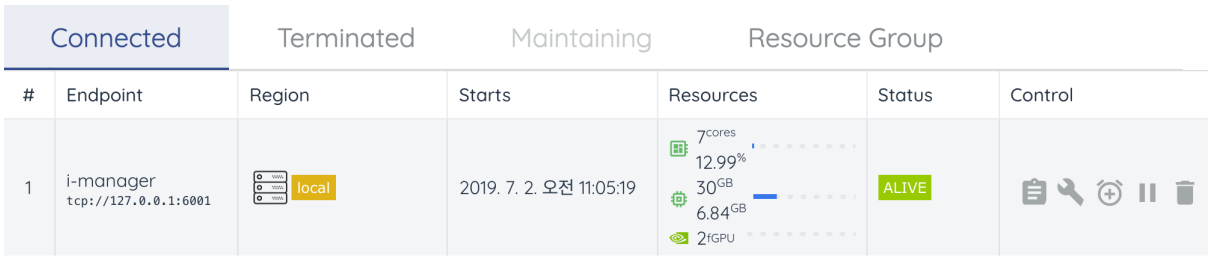


You can set resources such as CPU, RAM, fGPU, and etc. to be provided by the resource preset by clicking the settings icon (cogwheel) in the Controls column. In the example below, the GPU field is disabled since the GPU provision mode of the Backend.AI server is set to "fractional". After setting the resources with the desired values, save it and check if the corresponding preset is displayed when creating the compute session. If you can only allocate resources less than the amount of resources defined in the preset, the corresponding preset would not be printed.

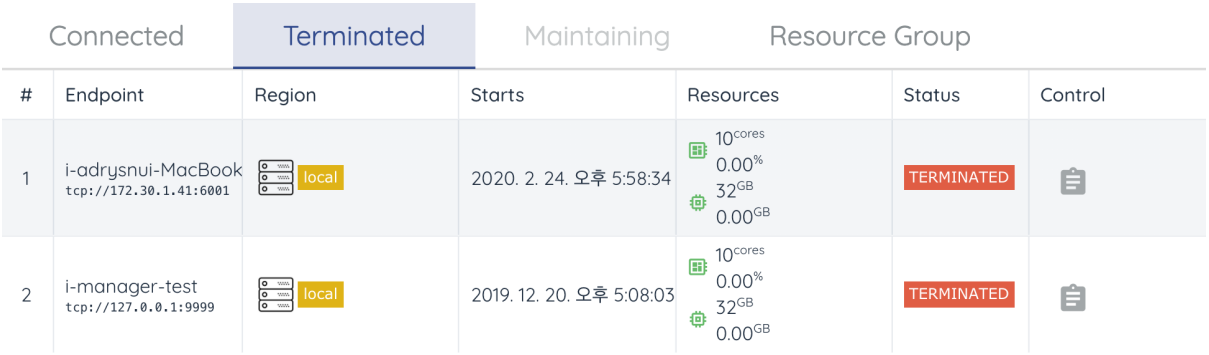


10.8 Query agent nodes

Superadmins can view the list of agent worker nodes, currently connected to Backend.AI, by visiting the Resources page. You can check agent node's IP, connecting time, actual resources currently in use, and etc. The Console does not provide the ability to manipulate agent nodes.



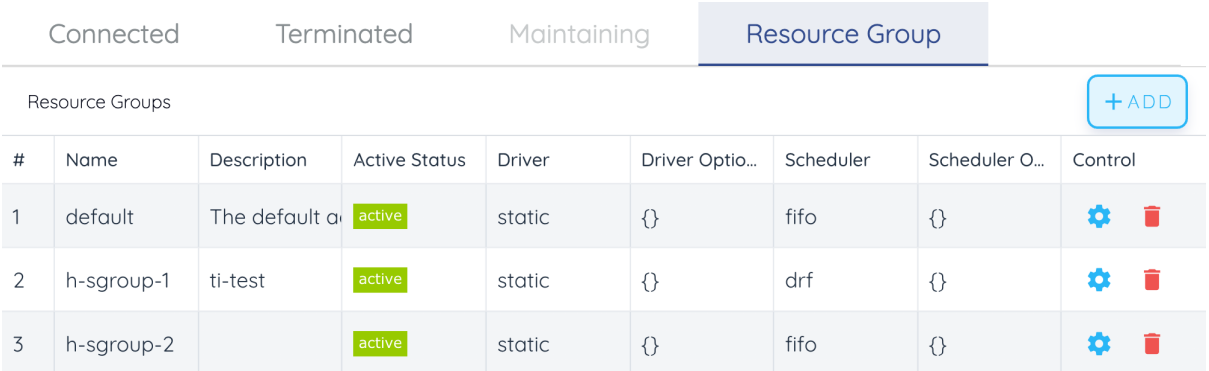
On Terminated tab, you can check the information of the agents that has been connected once and then terminated or disconnected. It can be used as a reference for node management.



10.9 Manage resource group

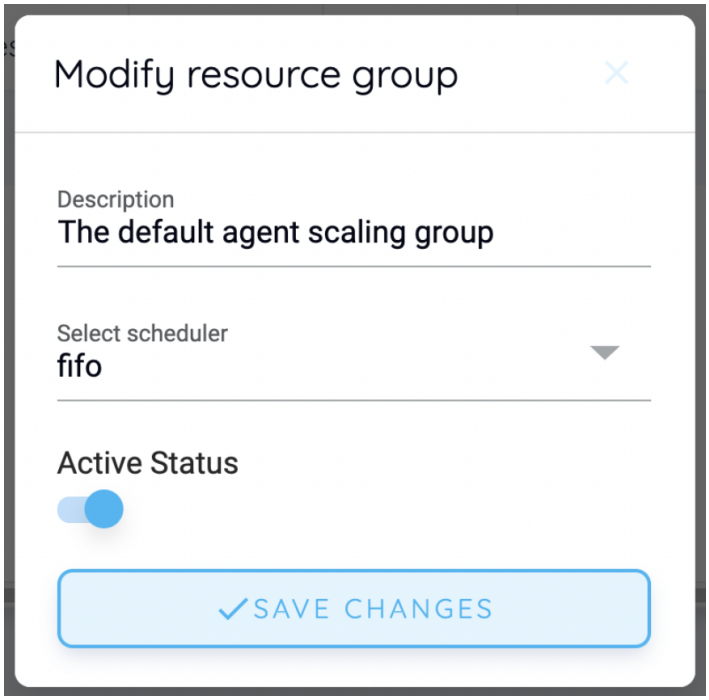
Agents can be grouped into units called resource (scaling) groups. For example, let's say there are 3 agents with V100 GPUs and 2 agents with P100 GPUs. You want to expose two types of GPUs to users separately, then you can group three V100 agents into one resource group, and the remaining two P100 agents into another resource group.

Adding a specific agent to a specific resource group is not currently handled in the UI, and it can be done by editing agent config file from the installation location and restart the agent daemon. Management of the resource groups is possible in Scaling Group tab of the Resource page.



You can edit a resource group by clicking the settings icon in the Control column. In the Select scheduler field, you can choose the scheduling method for creating a compute session. Currently, there are three types: FIFO, LIFO, and DRF. FIFO and LIFO are scheduling methods creating the first- or the last-enqueued compute session in the job queue. DRF stands for Dominant Resource Fairness, and it aims to provide resources

as faie as possible for each users. You can deactivate a resource policy by turning off Active Status.



You can create a new resource policy by clicking the CREATE button.

10.10 System settings

In the System Settings page, you can see main settings set in Backend.AI server. Currently, it provides several controls which can change settings and setting list features.

Note: We will continue to add broader range of setting controls.

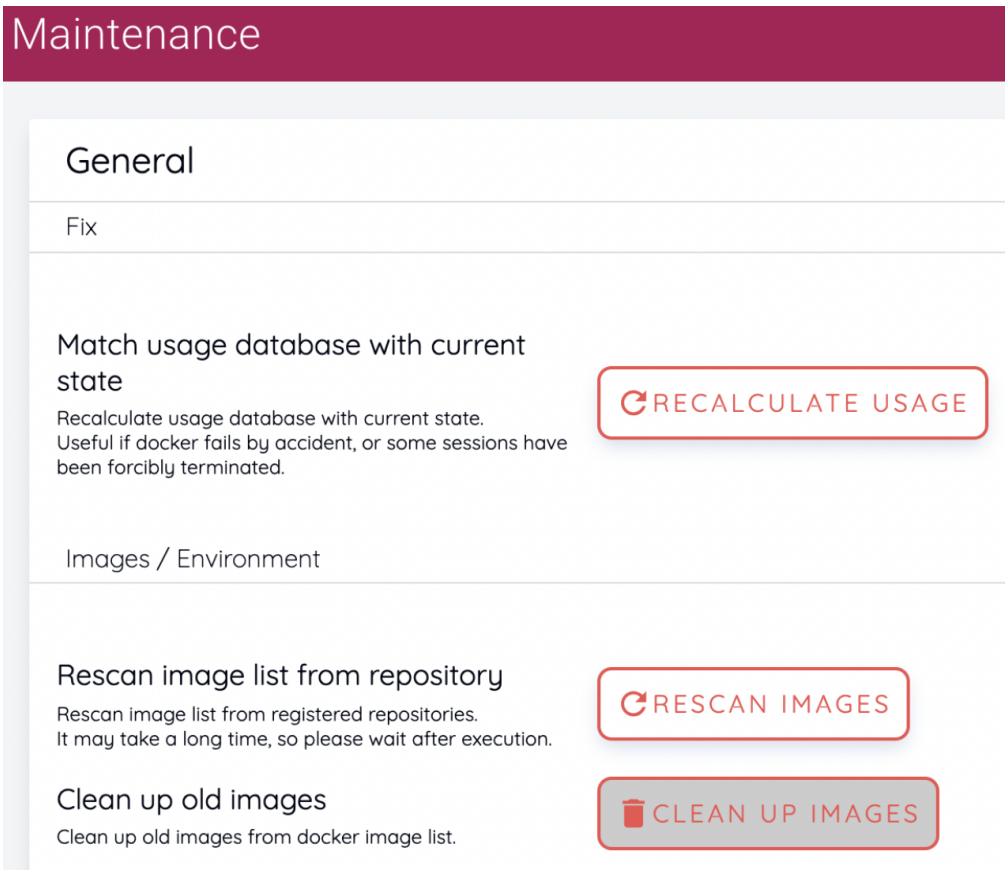
10.11 Server management

Go to the Maintenance page and you will see some buttons to manage the server.

- RECALCULATE USAGE: Occasionally, due to unstable network connections or

container management problem of Docker daemon, there may be a case where the resource occupied by Backend.AI does not match the resource actually used by the container. In this case, click the RECALCULATE USAGE button to manually correct the resource occupancy.

- RESCAN IMAGES: Update image meta information from all registered Docker registries. It can be used when a new image is pushed to a Backend.AI-connected docker registry.



Note: We will continue to add other settings needed for management, such as removing unused images or registering periodic maintenance schedules.

CHAPTER 11

Trouble Shooting

If you use the GUI Console for a long time, you may experience connection problems to Jupyter and/or terminal service, or compute session list not updating. Those problems often disappear when you refresh the Console page. You may refresh the Console by following methods.

- Web-based Console: Refresh the browser page (use the shortcut provided by browsers such as Ctrl-R). Since the browser's cache may cause troubles sometimes, it is recommended to refresh the page bypassing the cache (such as Shift-Ctrl-R, but the keys may different on each browser).
- Console App: Press Ctrl-R shortcut to refresh the app.

11.1 SFTP disconnection

When Console App launches SFTP connection, it uses a local proxy server which is embeded in the App. If you exit the Console App during the file transfer with SFTP

protocol, the transfer will immediately fail because the connection established through the local proxy server is disconnected. Therefore, even if you are not using a compute session, you should not quit the Console App while using SFTP. If you need to refresh the page, we recommend using the Ctrl-R shortcut.

Also, if the Console App is closed and restarted, the SFTP service is not automatically initiated for the existing compute session. You must explicitly start the SSH/SFTP service in the desired container to establish the SFTP connection.

11.2 Inconsistency between allocated and actually using resources

Note: This feature is only available for superadmins.

Occasionally, unstable network connection and container management issues by the docker daemon may cause incorrect display of allocated resources. If this problem occurs, you can go to the Maintenance page and click the RECALCULATE USAGE button to manually correct it.

11.3 Image is not displayed after it is pushed to a docker registry

Note: This feature is only available for superadmins.

If new image is pushed to one of Backend.AI docker registries, the image metadata must be updated in Backend.AI to be used in creating a compute session. Metadata update

can be performed by clicking the RESCAN IMAGES button on the Maintenance page. This will update metadata for every docker registries, if there are multiple registries.

If you want to update the metadata for a specific docker registry, you can go to the Registries tab in Environments page. Just click the refresh button in the Controls column of the desired registry. Be careful not to delete the registry by clicking the trash icon.

CHAPTER 12

Terms of License Agreement

12.1 Backend.AI License (Software)

This document defines the terms of the license agreement for the Backend.AI software. The usage fee and support plan of Backend.AI Cloud service provided by Lablup is independent of this policy.

Backend.AI server components (hereinafter referred to as “Backend.AI Server”) are distributed under the GNU Lesser General Public License v3.0 (“LGPL”), and API Client libraries and auxiliary components for accessing Backend.AI server (hereinafter “Backend.AI Client”) are distributed under the MIT License. Even if LGPL complies, commercial contracts with Lablup Co., Ltd. (“Lablup”) are required depending on the conditions when performing profit activities using the Backend.AI server. Several additional plug-ins and management Hubs targeting Backend.AI enterprise solutions are not open source, but commercial software.

Term Definition

- Hardware: Includes virtual machine and container environments where users have the right to run and software on their own or leased physical computers.
- Organization: Individuals, corporations, organizations, institutions (including non-profit and commerical organizations; however, subsidiaries that are separate corporations are not included)

LGPL must be followed when users use and change Backend.AI Server (Manager / Agent / Common) or develop and distribute software that uses it. Below is an example of when there is no obligation under the LGPL at this time:

1. In case of distributing software which imports it as a module (eg Python import) without changing the Backend.AI server.
2. When Backend.AI server is installed on the hardware and used by the general public over the network.

The correct interpretation of all other cases is subject to the LGPL original text and court judgment.

Apart from LGPL compliance, commercial contracts must be made with Lablup in the following cases:

1. When software that works only after installing the Backend.AI server is sold to customers outside the organization.
2. When selling hardware including Backend.AI server to customers outside the organization.
3. When the Backend.AI server is installed on the hardware and the usage fee is received from a customer outside the organization that uses it.

In other cases, you can use the Backend.AI server for free.

Interpretation Example

- If you distribute Backend.AI server with modifications to outside the organization, you must disclose the code and apply LGPL the same way. There is no obligation

to disclose the code if it is used internally only.

- Software using Backend.AI server as an essential library
  - Free distribution: The software does not have to be (L)GPL, and a separate contract with Lablup is not required.
  - Paid distribution: The software does not need to be (L)GPL, but a commercial contract with Lablup is required.
- Backend.AI server is installed on the hardware and it is distributed to the public
  - Free distribution: No separate contract with Lablup is required.
  - Paid distribution: A commercial contract with Lablup is required.
- Backend.AI server installed hardware
  - Free distribution: No separate contract with Lablup is required.
  - Paid distribution: A commercial contract with Lablup is required.

Commercial contracts include monthly / annual subscription fees for the enterprise version by default, but details may vary depending on individual contracts. Users of the open-source version can also purchase maintenance and support plans separately.

CHAPTER 13

References

- Online API Manual: https://docs.backend.ai
- Online Client SDK Manual: https://docs.client-py.backend.ai

The latest versions of this document can be found from sites below:

- https://console.docs.backend.ai/en/stable (English)
- https://console.docs.backend.ai/ko/stable (Korean)