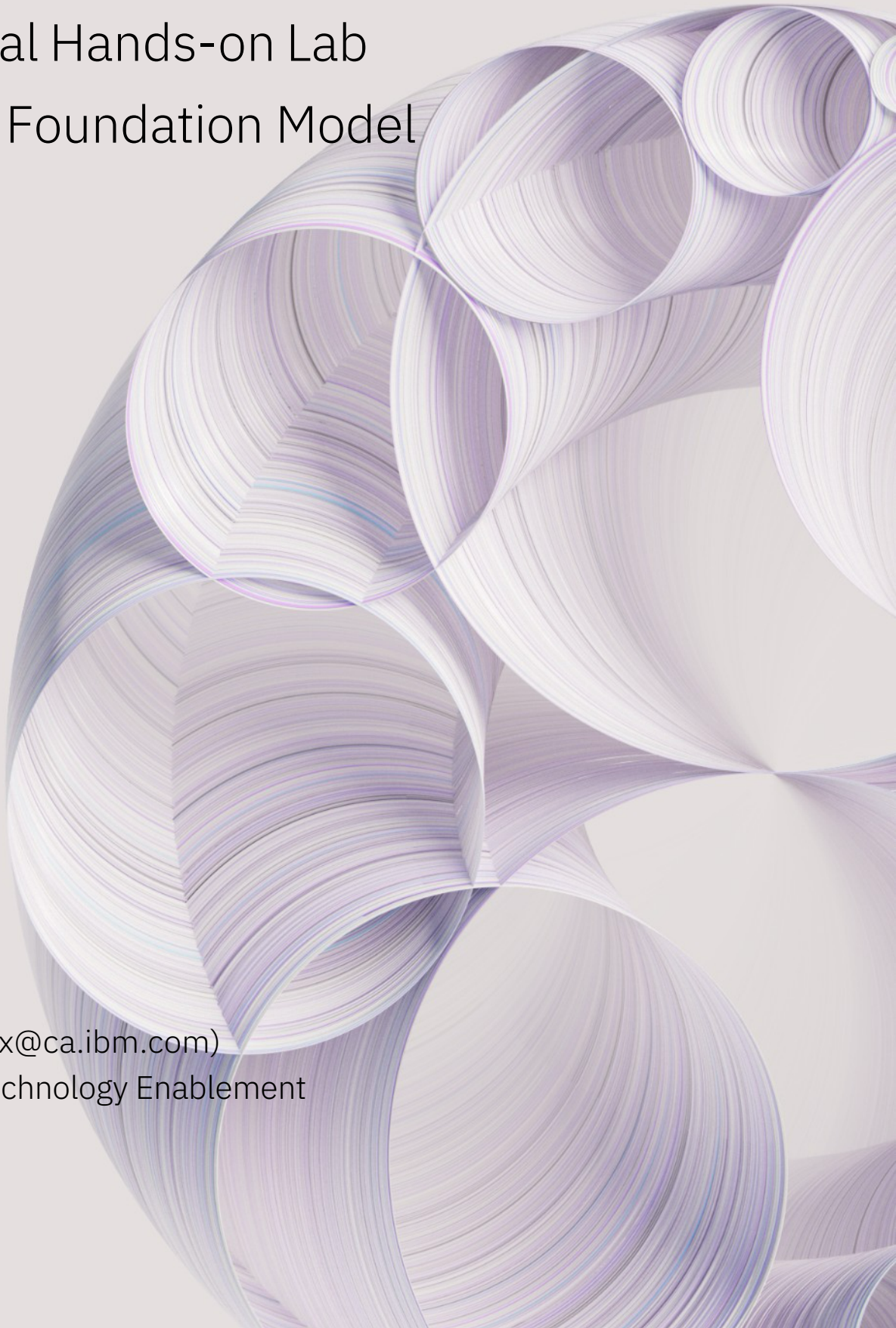
IBM watsonx.ai

Technical Hands-on Lab

Custom Foundation Model



Felix Lee (felix@ca.ibm.com)
Worldwide Technology Enablement

Contents

© 2024 IBM Corporation

1. Introducing watsonx.ai Custom Foundation Model

IBM's generative AI strategy is open, trusted, and performant. IBM watsonx.ai is designed to work with models from different sources: IBM-developed models, open source, and other third-party models.

IBM watsonx.ai provides a set of carefully curated models which includes IBM's Granite series models, Meta's llama models, Goggle's flan models, and Mistral AI's mistral and mixtral models. These models provide a wide range of generative AI use cases. For example:

- Generation
- Extraction
- Classification
- Summarization
- Question Answering
- Code generation and translation

IBM has recognized that clients may have been working with models that they have found satisfactory or might have tuned a model to support their business use cases. Clients will want to deploy these models within watsonx.ai to leverage the rich features. For example:

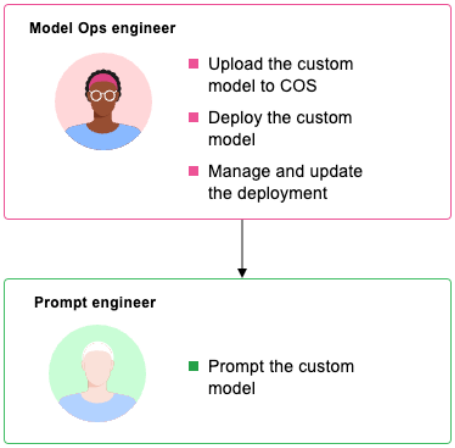
- Model governance, AI automation, and deployment
- Integration with watsonx.data and watsonx.governance
- Integration with various AI Assistants

The watsonx.ai Custom Foundation Model feature allows clients to bring their own model and deploy it on watsonx.ai. This feature provides the flexibility for clients to implement the AI solutions that are right for their use case.

© 2024 IBM Corporation

2. About this Lab

The process for deploying a foundation model and making it available for inferencing includes tasks that are performed by a Model Ops engineer and a Prompt Engineer. The following diagram shows a high-level flow of tasks typically performed by a Model Ops engineer and a Prompt engineer:



This lab focuses on the first 2 tasks of the Model Ops engineer: using the watsonx.ai Custom Foundation Model feature to upload a custom model to your Cloud Object Storage (COS) and then deploying a custom model to watsonx.ai.

The process covered in this lab applies to deploying a model in the watsonx.ai Software as a Service (SaaS) environment in the cloud. For deploying a custom model on-premises, consult the documentation at: https://tinyurl.com/br7mfnb2.

© 2024 IBM Corporation

2.1 Disclaimer

IBM watsonx.ai is developed and released in an agile manner. In addition to adding new capabilities, the web interface is likely to change over time. Therefore, the **screenshots used in this lab may not always look exactly like what you see**. You can expect to encounter some of the following:

- Additional foundation models in the library list.
- Changes in the user interface (location of buttons, text for various fields)

These should not affect how the labs work, but have patience and explore.

However, the list of model architectures supported for **Custom Foundation Model** may change. You should always consult **Supported model architectures** on Planning to deploy a custom foundation model to ensure the model you want to add is based on a supported architecture. Other prerequisites may also change (for example, the format of the model content).

Please raise questions on the #data-ai-demo-feedback Slack channel (IBMers only). IBM partners can request help at the Partner Plus Support website.

© 2024 IBM Corporation

3. Getting Help

Lab guide help: If you require assistance in interpreting any of the steps in this lab, please post your questions to the #data-ai-demo-feedback Slack channel (IBMers only). IBM partners can request help at the Partner Plus Support website.

IBM watsonx.ai: Assistance with the watsonx.ai product itself is available in the #watsonx-ai-feedback (IBMers) and the #watsonx-ai-enablement Slack channels (IBMers only). Additionally, please refer to the watsonx.ai documentation as needed.

© 2024 IBM Corporation

4. Prerequisites & Getting Started

- This lab downloads a model from Hugging Face. You will need to have a Hugging Face account to do so. You can go to https://huggingface.co/join and provide the necessary information to sign up for an account.

- You need a Mac OS or Linux environment with sufficient disk space (2 GB) to download the model used for this exercise from Hugging Face.

- **Python 3.9 or above** installed (more details below).

- This lab was verified using the **Chrome** browser. It does NOT work with **Safari.**

- You should be comfortable working with basic Linux/Unix navigation (switching directories, listing files, etc.).

- There are many steps in this lab. Your local workstation may also have existing Python environments. Be patient and be prepared to work through minor issues.

4.1 Examine your Python version

You must meet the minimum Python version requirements for this lab to work. Simply run the **python** command to verify what version you are using.

```
felixl@Felixs-MacBook-Pro ~ % python
Python 3.7.4 (main, Jul  5 2023, 08:40:20) [Clang 14.0.6 ] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

In the screen capture above, the Python version is 3.7.4, which is below the minimum requirement. This means that the lab will not work.

The process of updating Python on your workstation is beyond the scope of this lab. However, you will need to update to at least 3.9 for this lab to work. In this example, once Python is updated to 3.11, rerunning the **python** command shows the following:

```
felixl@Felixs-MacBook-Pro ~ % python
Python 3.11.4 (main, Jul  5 2023, 08:40:20) [Clang 14.0.6 ] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

The **python** command shows that the version is > 3.9. It is fine to proceed with the lab.

© 2024 IBM Corporation

4.1 Using TechZone

- **You must use a TechZone account** to perform this lab, or you will run into missing service or out-of-tokens issues. When you use TechZone, you **will not (and should not)** need to subscribe to any additional services.

- **Your TechZone account must support the Standard Plan** for Watson Machine Learning. This is a higher requirement than most TechZone patterns. You must use the specific pattern created for this lab (outlined in Section 5.4). This lab will NOT work if you reserve the standard watsonx.ai/.governance SaaS pattern.

© 2024 IBM Corporation

5. Custom Foundation Model

At a high level, adding your custom foundation model to watsonx.ai involves these steps:

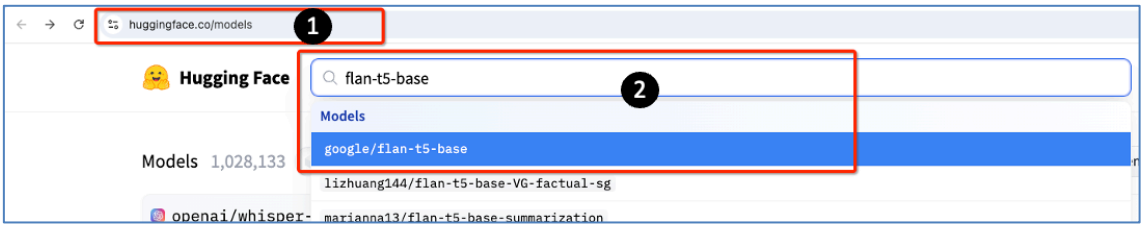
1. Verifying that watsonx.ai supports the model you want to add
2. Getting the model files (in this lab, you will get it from Hugging Face)
3. Converting the model file into the desired format (if necessary)
4. Creating a project (a sandbox in this exercise) in watsonx.ai (to do so requires making a specific TechZone reservation)
5. Storing the model files in a supported Cloud Object Storage (COS) using the storage bucket associated with your project
6. Creating a deployment space and setting up a connection to your COS instance.
7. Creating a model asset from the COS content in the deployment space
8. Deploying the model asset
9. Using the Custom Foundation Model

5.1 Verifying your model is supported

First, you need to ensure the model you want to use is supported by watsonx.ai. For this lab, you will use the **flan-t5-base** model as an example. The purpose of this lab is to demonstrate the process. The steps are identical for other models.

1. Go to https://huggingface.co/models

2. Enter **flan-t5-base** in the search field. Hugging Face will pull up the relevant entry.

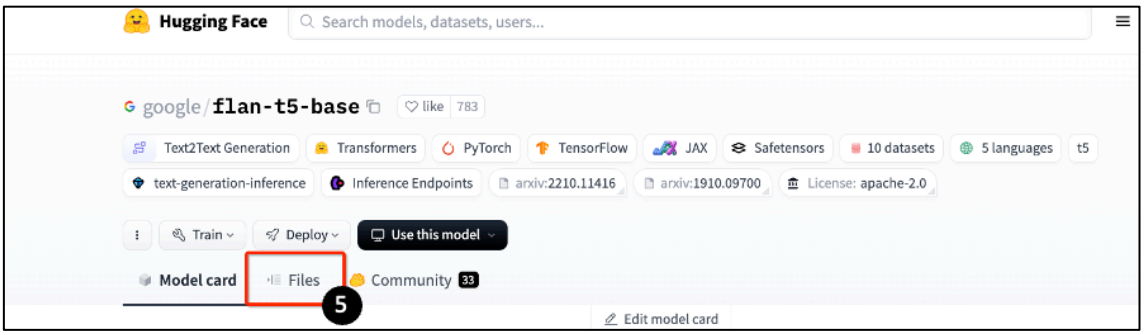


3. Select **google/flan-t5-base** and its associated Hugging Face page appears, displaying the model card for the **flan-t5-base** model.

   Note: **google/flan-t5-base** is the model name (including the "google" part).

4. Click the **Files** tab.

© 2024 IBM Corporation



5. The **Files** tab opens. For this lab, you require the following entries (or the model cannot be deployed to watsonx.ai):

    a. **config.json** - required to load the model in the **Text Generation Inference Service** (TGIS) runtime.

    b. **tokenizer.json**

    c. **model.safetensors** (the model must be in this format)

6. Click on the **config.json** file and open it. Examine the content and look for **model_type** which indicates the model family/architecture. You must ensure that this belongs to the list of supported architecture.

    In this case, the **model_type** is **t5**.

7. Always check the model against the list of supported architecture. The model you want to deploy to watsonx.ai must have a supported architecture. As of November 2024, the list reads:

© 2024 IBM Corporation

Supported model architectures, quantization methods, parallel tensors support, and configuration sizes

Model architecture type	Supported quantization method	Supports parallel tensors (multiGpu)	Supported configurations
`bloom`	N/A	Yes	Small, Medium, Large
`codegen`	N/A	No	Small
`falcon`	N/A	Yes	Small, Medium, Large
`gpt_bigcode`	`gptq`	Yes	Small, Medium, Large
`gpt-neox`	N/A	Yes	Small, Medium, Large
`gptj`	N/A	No	Small
`llama`	`gptq`	Yes	Small, Medium, Large
`mixtral`	`gptq`	No	Small
`mistral`	N/A	No	Small
`mt5`	N/A	No	Small
`mpt`	N/A	No	Small
`t5`	N/A	Yes	Small, Medium, Large

As shown above, **t5** is a supported architecture with a Small, Medium, or Large watsonx.ai configuration (you will use this information later).

Now that you have verified the model you will use in this lab is supported, you need to download the model files from Hugging Face.

© 2024 IBM Corporation

5.2 Downloading the asset from Hugging Face

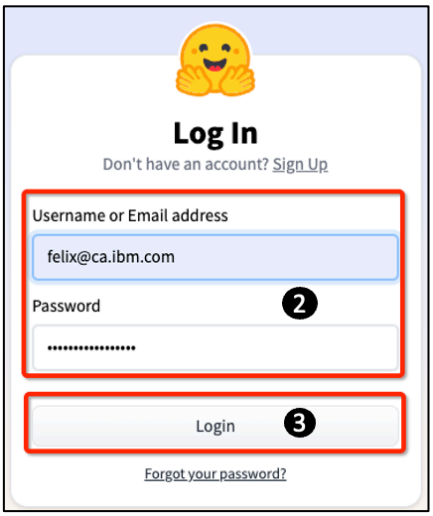
You need to have a Hugging Face account to perform the instructions in this section. If you do not already have one, you can go to https://huggingface.co/join and provide the necessary information to sign up.

5.2.1 Getting a Hugging Face Access Token

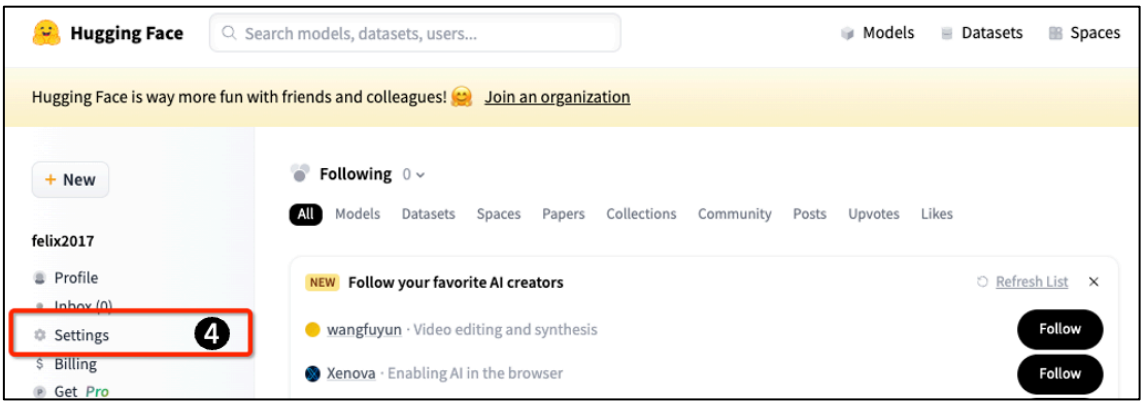
1. You need a Hugging Face access token to download model files from Hugging Face. Go to https://huggingface.co/ and click **Log in** to log into your Hugging Face account.

2. Enter your **Username/email**, **Password**.

3. Click **Login**



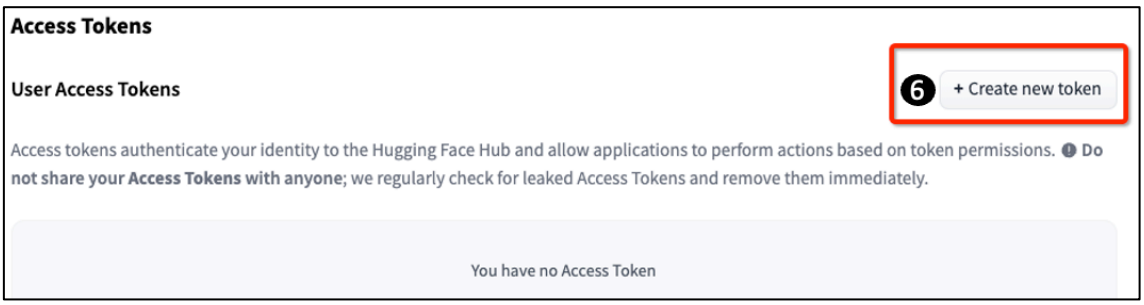
4. Click **Settings** on the left.



5. Click **Access Tokens** on the left.

6. Click **+ Create new token**.

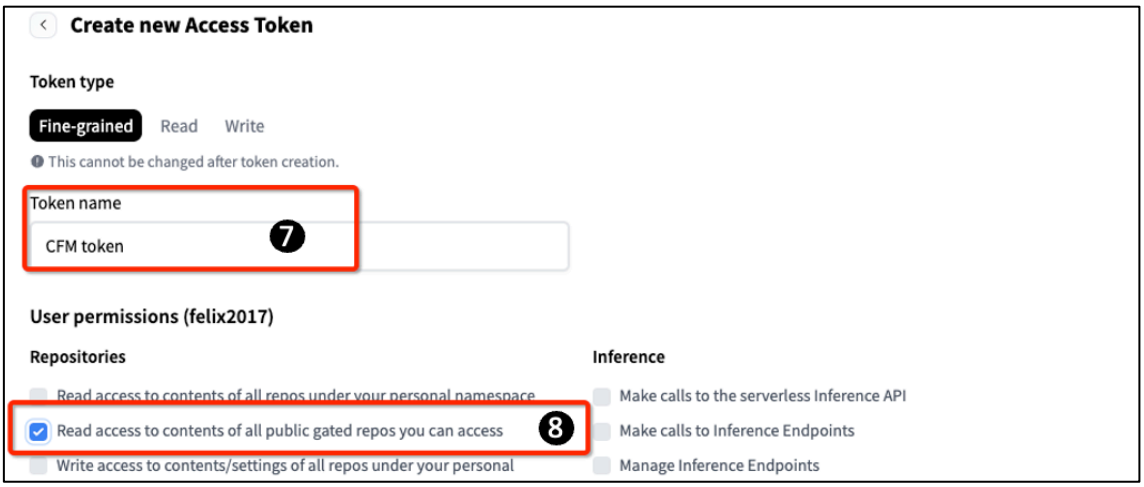
   Note: You may already have an access token. However, you may still want to create a new one for this lab.



7. Enter **CFM token** in the **Token name** field.

8. Select **Read access to contents of all public gated repos you can access** (you can select other permissions as well, but use this permission for the lab).

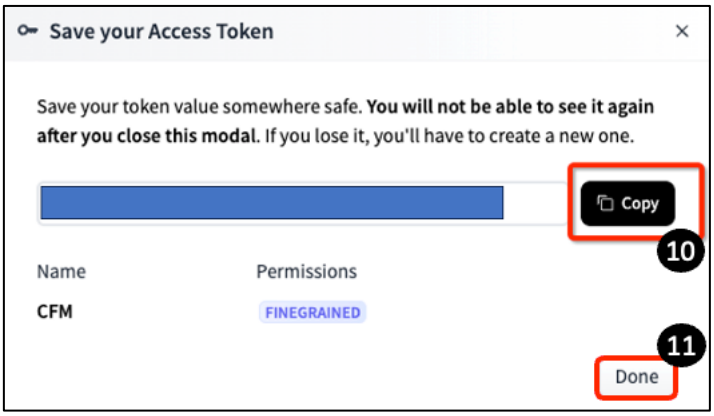
© 2024 IBM Corporation



9. Scroll down to the bottom and click **Create token**

10. The **Save your Access Token** page opens. As you can see below, the actual token is intentionally redacted in this lab for obvious reasons. Click **Copy** and save the tokens somewhere safe where you will be able to find and copy. Once you close this window, you will not be able to retrieve this token. If you did not save it, you will have to create another one.

11. Click **Done**.



Next, you need to download the **flan-t5-base** model from Hugging Face

© 2024 IBM Corporation

5.2.2 Downloading a model from Hugging Face

1. On your Mac OS or Linux workstation, open a terminal session.

2. You should create a special subdirectory and create a new Python environment. In this example, a subdirectory **CFM** is created under the **/Users/felixl/Downloads** directory. You will substitute in your appropriate values (for example, **/home/<username>**, or /**home/<username>/Downloads**); where you use your name as opposed to **felixl**.

   Enter these commands:

   cd /Users/<username>/Downloads
   mkdir CFM
   python -m venv /Users/<username>/Downloads/CFM/venv
   ls /Users/<username>/Downloads/CFM

   You should now see a **venv** directory under the **CFM** directory.

```
felixl@Felixs-MacBook-Pro ~ % cd /Users/felixl/Downloads
felixl@Felixs-MacBook-Pro Downloads % python -m venv /Users/felixl/Downloads/CFM/venv
felixl@Felixs-MacBook-Pro Downloads % ls /Users/felixl/Downloads/CFM
venv
felixl@Felixs-MacBook-Pro Downloads %
```

3. Use the newly created Python environment by issuing the following commands:

   cd /Users/<username>/Downloads/CFM
   source venv/bin/activate

   Note how the left part of the terminal changes to reflect you are in the virtual environment **(venv)**.

```
felixl@Felixs-MacBook-Pro Downloads % cd /Users/felixl/Downloads/CFM
felixl@Felixs-MacBook-Pro CFM % source venv/bin/activate
(venv) felixl@Felixs-MacBook-Pro CFM %
(venv) felixl@Felixs-MacBook-Pro CFM %
```

4. Create a subdirectory under the **CFM** directory to host the model you are going to download by entering the following commands:

   mkdir flan-t5-base
   ls -l

   You should see this:

© 2024 IBM Corporation

```
(venv) felixl@Felixs-MacBook-Pro CFM % mkdir flan-t5-base
(venv) felixl@Felixs-MacBook-Pro CFM % ls -l
total 0
drwxr-xr-x  2 felixl  staff   64  3 Oct 12:51 flan-t5-base
drwxr-xr-x  6 felixl  staff  192  3 Oct 12:30 venv
(venv) felixl@Felixs-MacBook-Pro CFM %
```

5. Install the **huggingface-cli** package using the following command:

   pip install -U “huggingface_hub[cli]”

   You will see something similar to this:

```
(venv) felixl@Felixs-MacBook-Pro CFM % pip install -U "huggingface-hub[cli]"
Requirement already satisfied: huggingface-hub[cli] in /Users/felixl/miniconda3/lib/python3.11
/site-packages (0.23.0)
Collecting huggingface-hub[cli]
  Using cached huggingface_hub-0.24.6-py3-none-any.whl (417 kB)
Requirement already satisfied: filelock in /Users/felixl/miniconda3/lib/python3.11/site-packag
es (from huggingface-hub[cli]) (3.14.0)
Requirement already satisfied: fsspec>=2023.5.0 in /Users/felixl/miniconda3/lib/python3.11/sit
e-packages (from huggingface-hub[cli]) (2024.3.1)
Requirement already satisfied: packaging>=20.9 in /Users/felixl/miniconda3/lib/python3.11/site
-packages (from huggingface-hub[cli]) (23.2)
Requirement already satisfied: pyyaml>=5.1 in /Users/felixl/miniconda3/lib/python3.11/site-pac
kages (from huggingface-hub[cli]) (6.0.1)
Requirement already satisfied: requests in /Users/felixl/miniconda3/lib/python3.11/site-packag
es (from huggingface-hub[cli]) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in /Users/felixl/miniconda3/lib/python3.11/site-pa
ckages (from huggingface-hub[cli]) (4.65.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /Users/felixl/miniconda3/lib/pyth
on3.11/site-packages (from huggingface-hub[cli]) (4.11.0)
Collecting InquirerPy==0.3.4 (from huggingface-hub[cli])
  Using cached InquirerPy-0.3.4-py3-none-any.whl (67 kB)
Collecting pfzy<0.4.0,>=0.3.1 (from InquirerPy==0.3.4->huggingface-hub[cli])
  Using cached pfzy-0.3.4-py3-none-any.whl (8.5 kB)
Requirement already satisfied: prompt-toolkit<4.0.0,>=3.0.1 in /Users/felixl/miniconda3/lib/py
thon3.11/site-packages (from InquirerPy==0.3.4->huggingface-hub[cli]) (3.0.39)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/felixl/miniconda3/lib/python
3.11/site-packages (from requests->huggingface-hub[cli]) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /Users/felixl/miniconda3/lib/python3.11/site-pa
ckages (from requests->huggingface-hub[cli]) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/felixl/miniconda3/lib/python3.11/s
ite-packages (from requests->huggingface-hub[cli]) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in /Users/felixl/miniconda3/lib/python3.11/s
ite-packages (from requests->huggingface-hub[cli]) (2023.5.7)
Requirement already satisfied: wcwidth in /Users/felixl/miniconda3/lib/python3.11/site-package
s (from prompt-toolkit<4.0.0,>=3.0.1->InquirerPy==0.3.4->huggingface-hub[cli]) (0.2.6)
Installing collected packages: pfzy, InquirerPy, huggingface-hub
  Attempting uninstall: huggingface-hub
    Found existing installation: huggingface-hub 0.23.0
    Not uninstalling huggingface-hub at /Users/felixl/miniconda3/lib/python3.11/site-packages,
 outside environment /Users/felixl/Downloads/CFM/venv
    Can't uninstall 'huggingface-hub'. No files were found to uninstall.
Successfully installed InquirerPy-0.3.4 huggingface-hub-0.24.6 pfzy-0.3.4
(venv) felixl@Felixs-MacBook-Pro CFM %
```

   Note: If your Python version is below 3.9, you may not have the pip command available in your venv. If this is an issue, check your Python version and upgrade it before proceeding.

6. Verify that **huggingface-cli** is set up by running the following command:

   huggingface-cli --help

You should see help output for this command as follows:

```
(venv) felixl@Felixs-MacBook-Pro CFM % huggingface-cli --help
usage: huggingface-cli <command> [<args>]

positional arguments:
  {env,login,whoami,logout,repo,upload,download,lfs-enable-largefiles,lfs-multipart-upload,sca
n-cache,delete-cache,tag}
                        huggingface-cli command helpers
    env                 Print information about the environment.
    login               Log in using a token from huggingface.co/settings/tokens
    whoami              Find out which huggingface.co account you are logged in as.
    logout              Log out
    repo                {create} Commands to interact with your huggingface.co repos.
    upload              Upload a file or a folder to a repo on the Hub
    download            Download files from the Hub
    lfs-enable-largefiles
                        Configure your repository to enable upload of files > 5GB.
    scan-cache          Scan cache directory.
    delete-cache        Delete revisions from the cache directory.
    tag                 (create, list, delete) tags for a repo in the hub

options:
  -h, --help            show this help message and exit
(venv) felixl@Felixs-MacBook-Pro CFM %
```

Note: Hugging Face CLI will deploy a version based on your Python version. If you run the huggingface-cli –help command and you do NOT see a download option, it is indicative that you have the wrong Python version. Check and upgrade your Python before proceeding.

7. You will need to use the Hugging Face token you saved in Section 5.2.1 Step 10. But first, you need to set up a few environment variables by entering the several commands in the terminal as follows:

   export HF_TOKEN="<your Hugging Face token>" (intentionally redacted below)
   export MODEL_NAME="google/flan-t5-base”
   export MODEL_DIR="<your base directory>/CFM/flan-t5-base"

   Note: you need to use the full path name you saw in Section 5.1, Step 3: google/flan-t5-base.

   For this example, the last variable is set to /Users/felixl/Downloads/CFM/flan-t5-base. (the token is redacted in the example below).

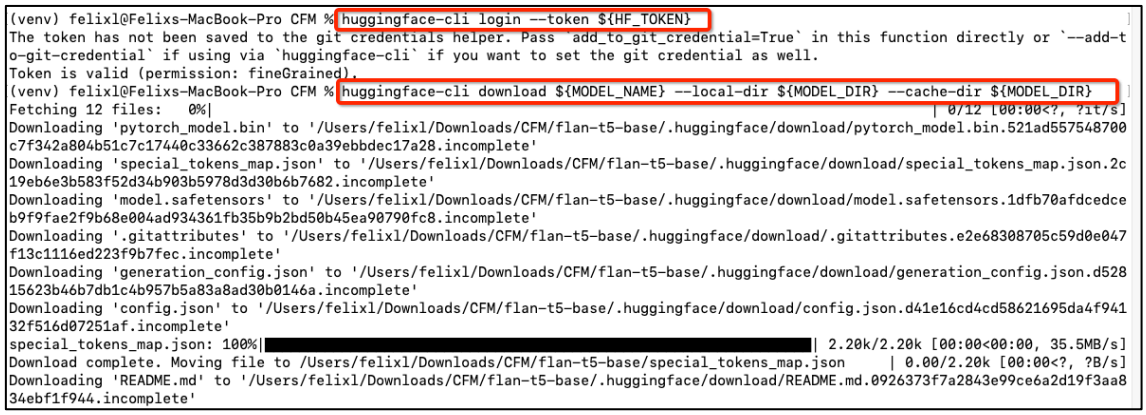
```
(venv) felixl@Felixs-MacBook-Pro CFM % export HF_TOKEN="[redacted]"
(venv) felixl@Felixs-MacBook-Pro CFM % export MODEL_NAME="google/flan-t5-base"
(venv) felixl@Felixs-MacBook-Pro CFM % export MODEL_DIR="/Users/felixl/Downloads/CFM/flan-t5-base"
(venv) felixl@Felixs-MacBook-Pro CFM %
```

   On your terminal, log into Hugging Face and download the model with these commands:

   huggingface-cli login --token ${HF_TOKEN}
   huggingface-cli download ${MODEL_NAME} --local-dir ${MODEL_DIR} --cache-dir ${MODEL_DIR}

© 2024 IBM Corporation

You should see something like this (the output is partially captured here):



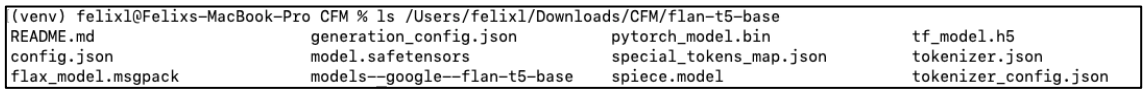
A total of 12 files should be downloaded.

Note: the **flan-t5-base** model is relatively small (about 4 GB in size). Be aware that some models are huge (think hundreds of GBs).

8. Verify the download with the command (sub in your appropriate directory):

   ls /Users/felixl/Downloads/CFM/flan-t5-base

   You should see this:



5.3 Converting the model to the desired format (if necessary)

To deploy a custom foundation model on watsonx.ai, that models need to be in **safetensors** format. In the example here, the **flan-t5-base** model artifacts include a **model.safesensors** file. If the model you downloaded does not have this format, you will need to convert the model to **safetensors** format.

The instructions to do so are documented in [Downloading a custom foundation model and setting up Storage] - look under the section titled **Converting a model to the required format**.

© 2024 IBM Corporation

5.4 Making a TechZone reservation and uploading model files

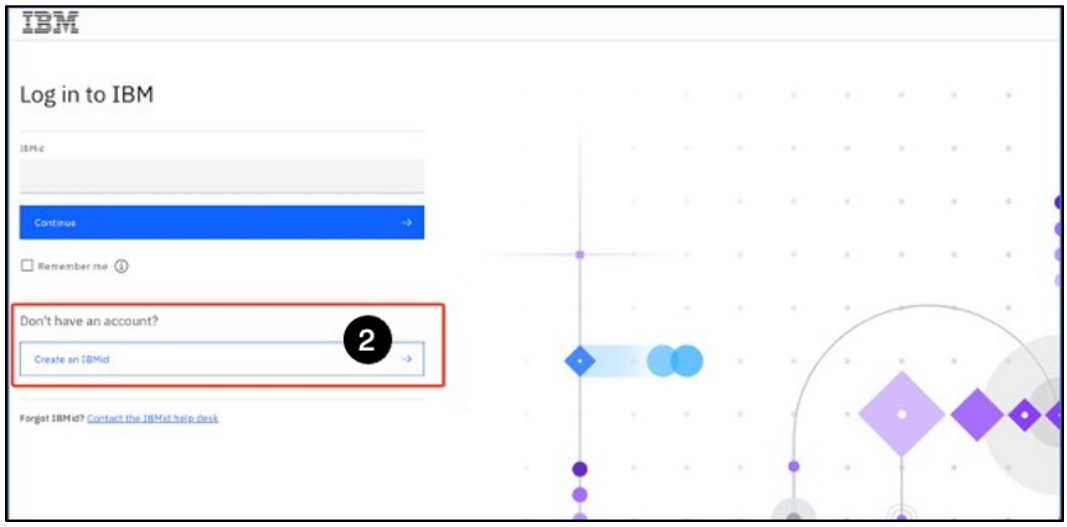
Once you have downloaded the model files, you are ready to set up watsonx.ai. As mentioned in Section 4, using **Custom Foundation Model** requires a specific TechZone pattern. In this section, you will make a reservation here and then upload the model files you just downloaded into the Cloud Object Storage (COS).

5.4.1 Reserving a specific TechZone pattern

The detailed setup instructions are as follows.

1. Go to the IBM TechZone website.

2. If you already have an IBMid, skip to Step 3. If you do not have one, click **Create an IBMid**.

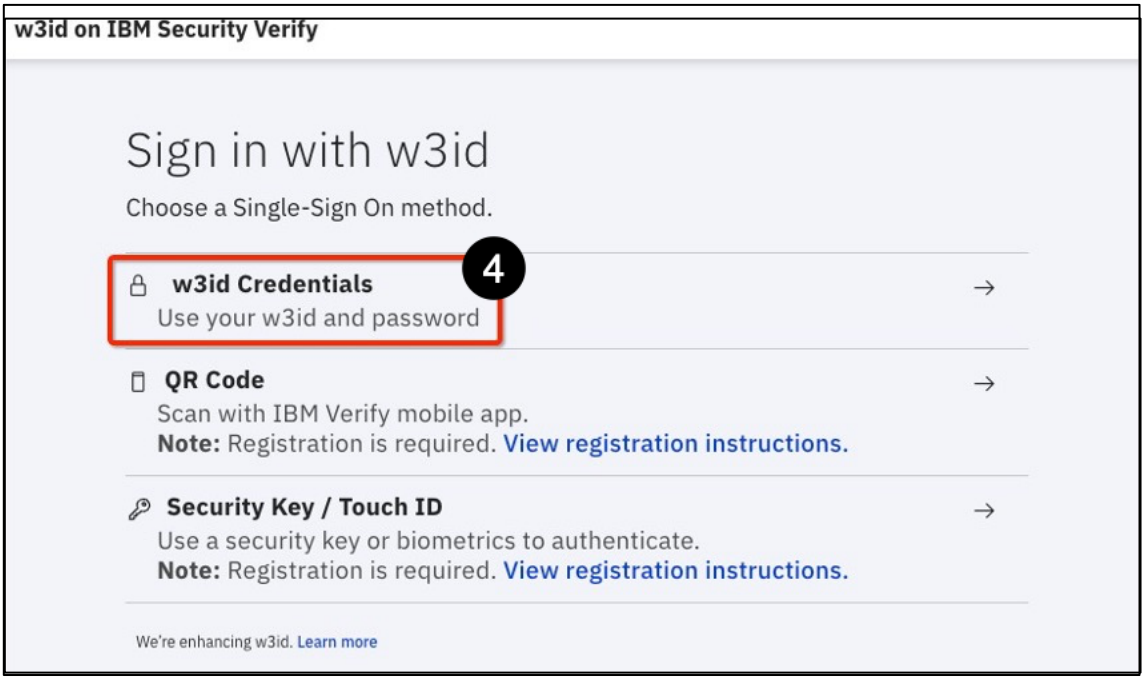


   Simply follow the instructions from the subsequent panels to create your IBMid.

3. Provide your IBMid and click **Continue**.

© 2024 IBM Corporation

4. Select your Single-Sign-On method. The **w3id Credentials** option is selected in this example:



5. Provide the necessary information (w3id and password, QR code, etc.) and sign in.

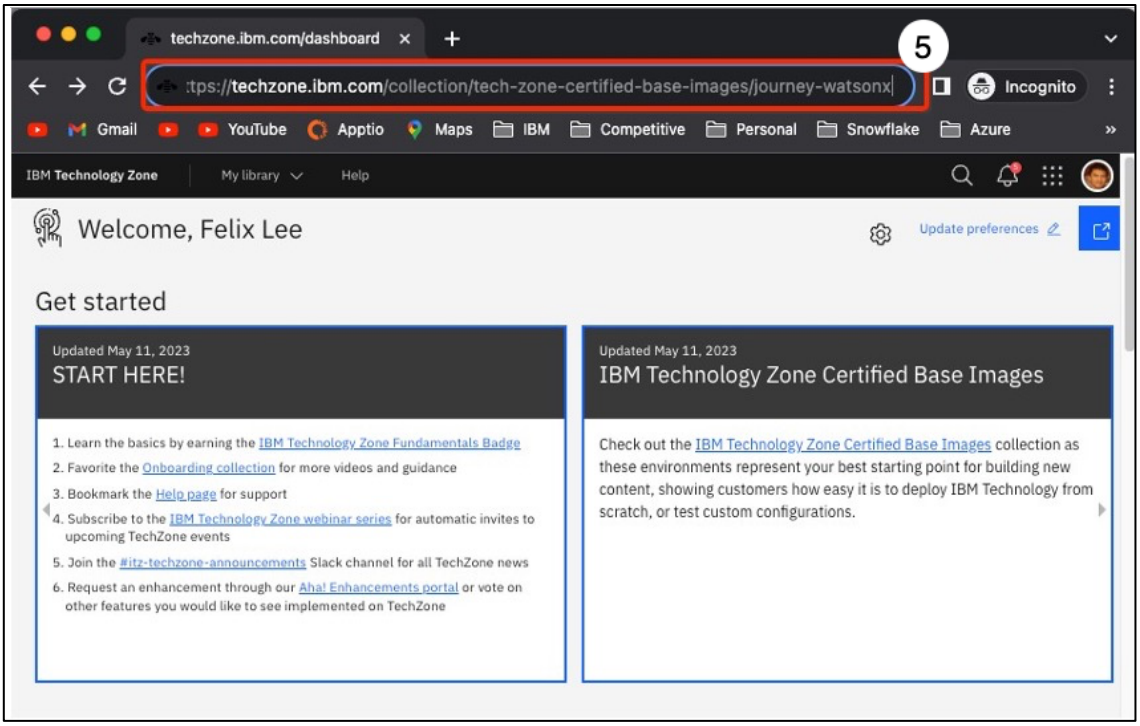
   If this is the very first time you are logging onto the TechZone, you will be brought to the TechZone dashboard. Copy and paste the following HTTP address in the browser's search field (do not use the magnifying glass icon) and press **Enter**:

   https://techzone.ibm.com/collection/tech-zone-certified-base-images/journey-watsonx

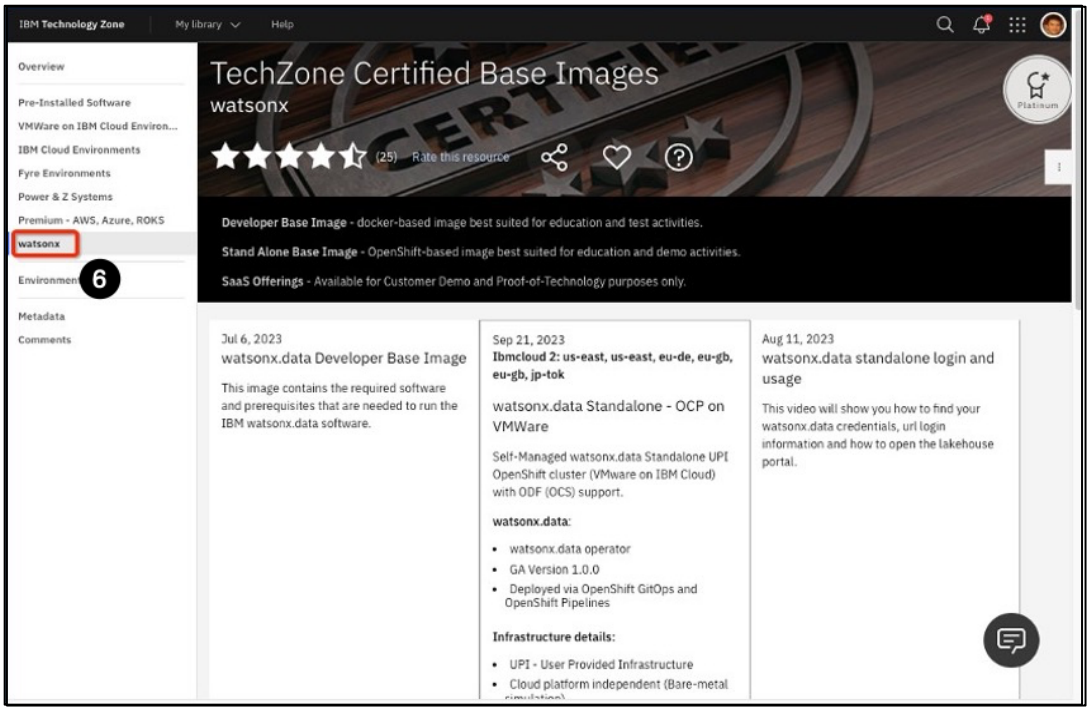
   This will take you to the **TechZone Certified Base Images** page.

   If you have been to TechZone before, you might be taken directly to the TechZone Certified Base Images page already. In that case, proceed to Step 6.

© 2024 IBM Corporation

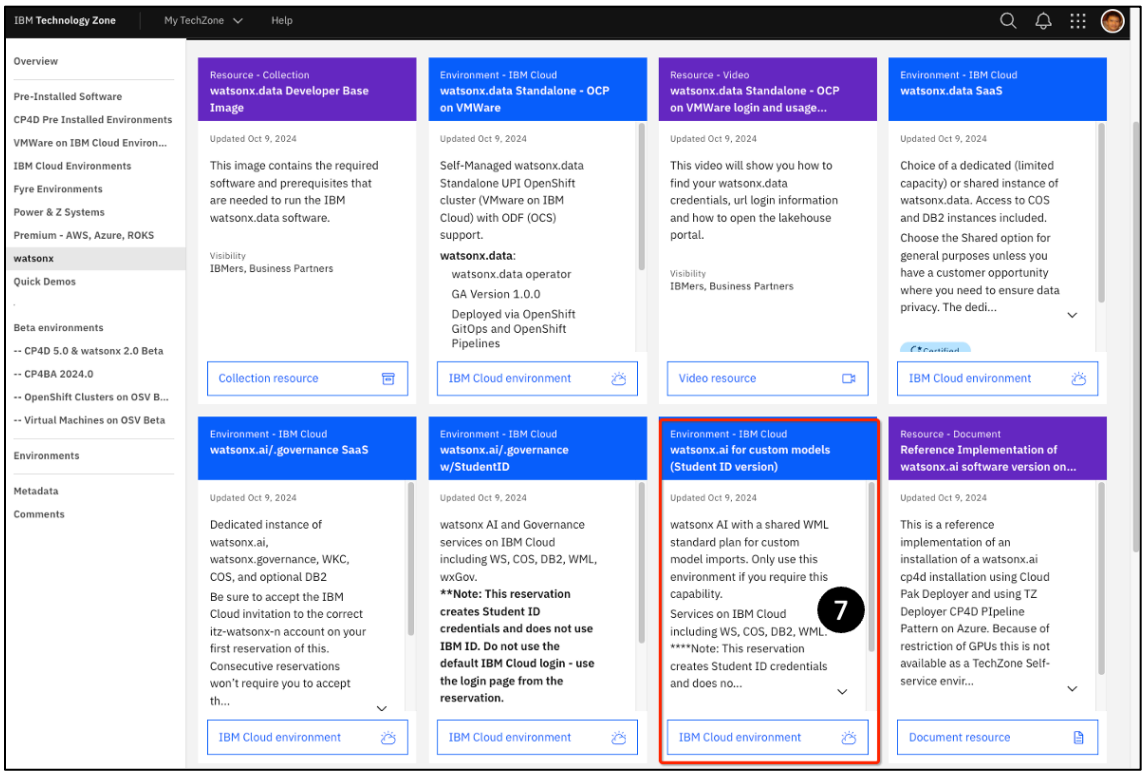


6. Your browser will open to the TechZone Certified Base Images page. On the left-hand side, the **watsonx** section should be highlighted. If not, click on it.

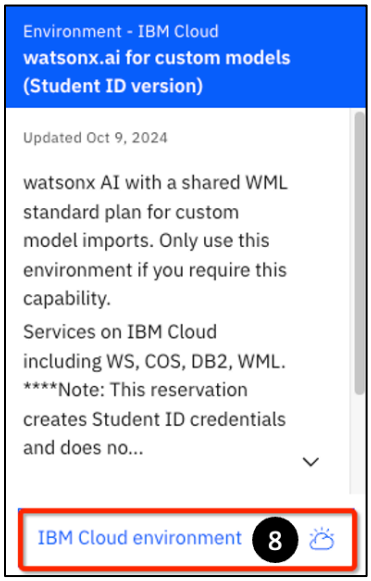


7. Scroll down until you find the **watsonx.ai for custom models (Student ID version)** tile (highlighted in red below). You **MUST** use this pattern.

© 2024 IBM Corporation

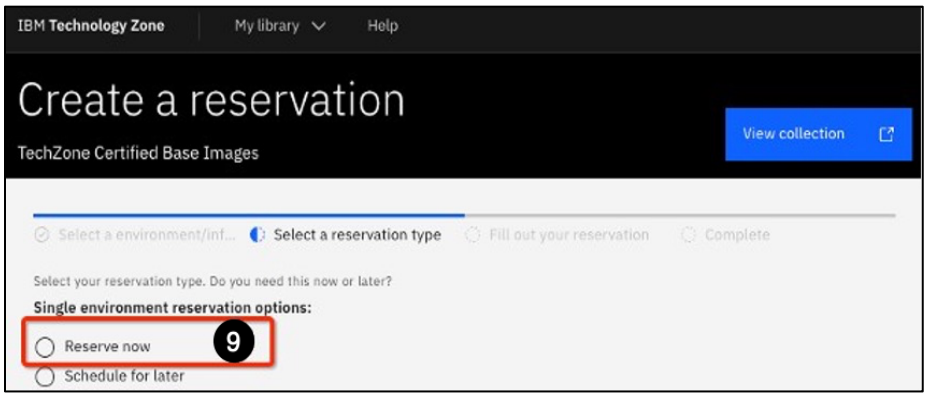


8. Hover over the **IBM Cloud environment Reserve** button at the bottom and the text will change to **Reserve it**. Click **Reserve it**.



9. On the **Create a reservation** page, select **Reserve now**. If you are not ready to use the environment in the next hour or so, you can select the **Schedule for later** radio button (the steps that follow are the same).

© 2024 IBM Corporation

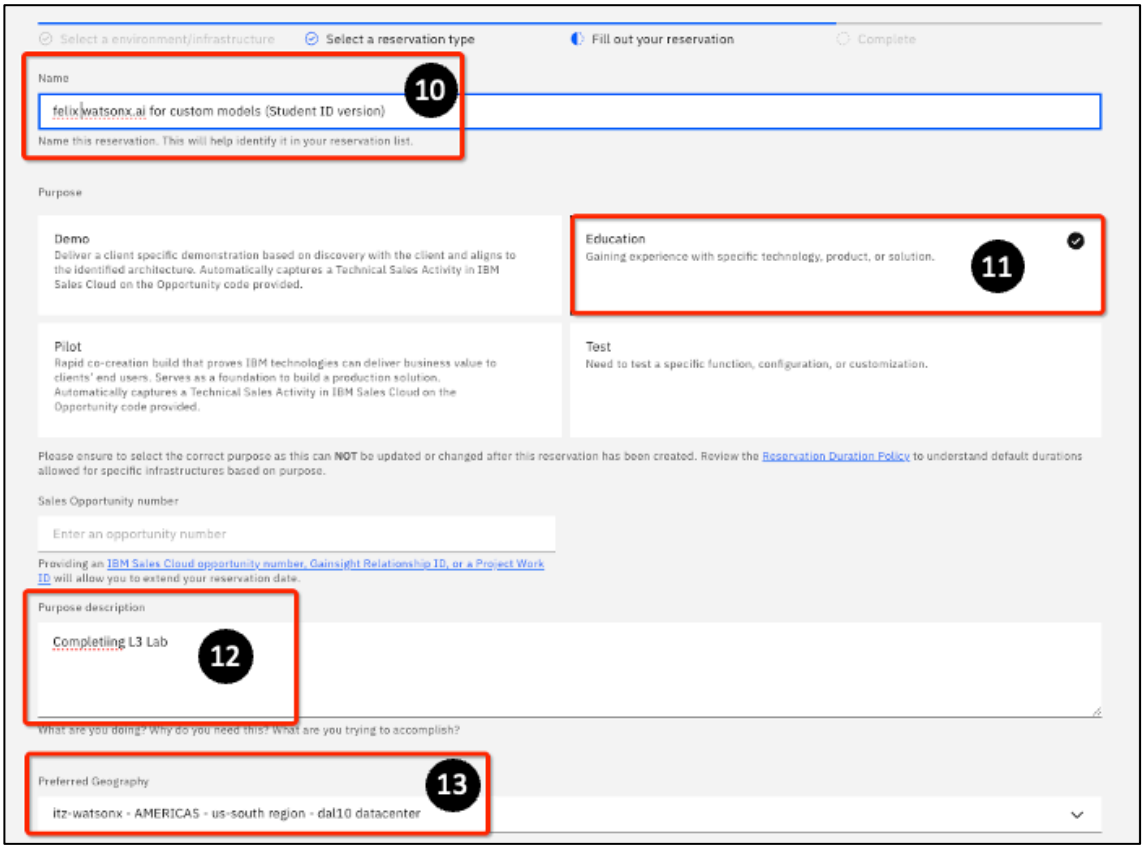


10. You can leave the **Name** as is, or change the **Name** of the reservation to **<your id> watsonx.ai for custom models (Student ID version).**

11. Click the **Education** tile.

12. In the **Purpose description** field, enter **Completing L3 Lab.**

13. Select your **preferred Geography**. You can choose any one you wish. In the example here **AMERICAS** is chosen.



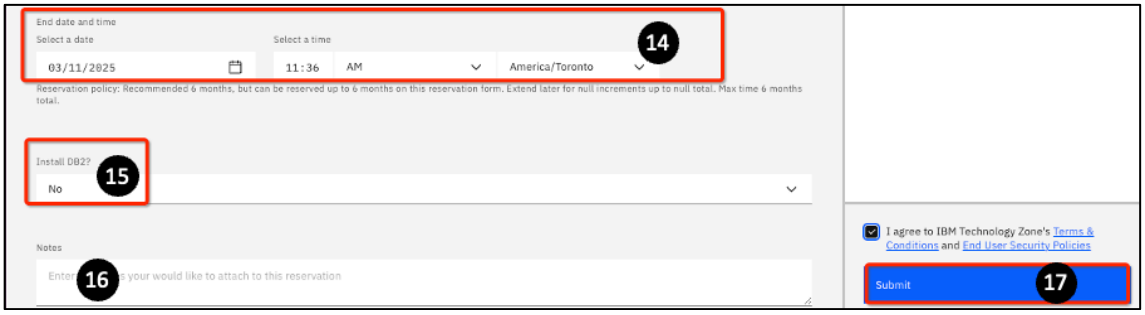
© 2024 IBM Corporation

14. Leave the **End date and time** as is.

15. Select **No** (should be the default) for installing Db2.

16. You can optionally provide some additional details in the **Notes** field if you wish.

17. On the right-hand side at the bottom, accept the **Terms and Conditions**, then click **Submit**.



18. You will get an email sent to your IBMid email address letting you know that your reservation is being provisioned. Once your watsonx.ai instance provisioning is completed, you will receive a second email telling you that it is ready for use.

    Note: Unlike other TechZone reservations, you do NOT need to accept an invitation to work with these accounts. You will be logging in via a "student login" described in the next section.

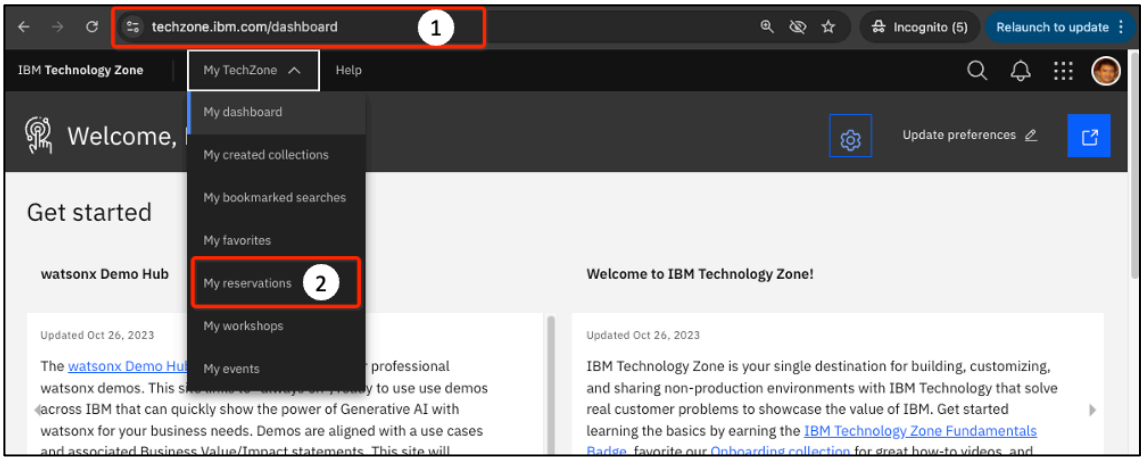
© 2024 IBM Corporation

5.4.2 Logging into your TechZone environment

To log into the **Custom Foundation Model** TechZone environment, you require information available on your reservation. You **DO NOT** log in using your IBM ID.

1. Go to http://techzone.ibm.com/dashboard

2. Click **My TechZone**, then click **My reservations**.

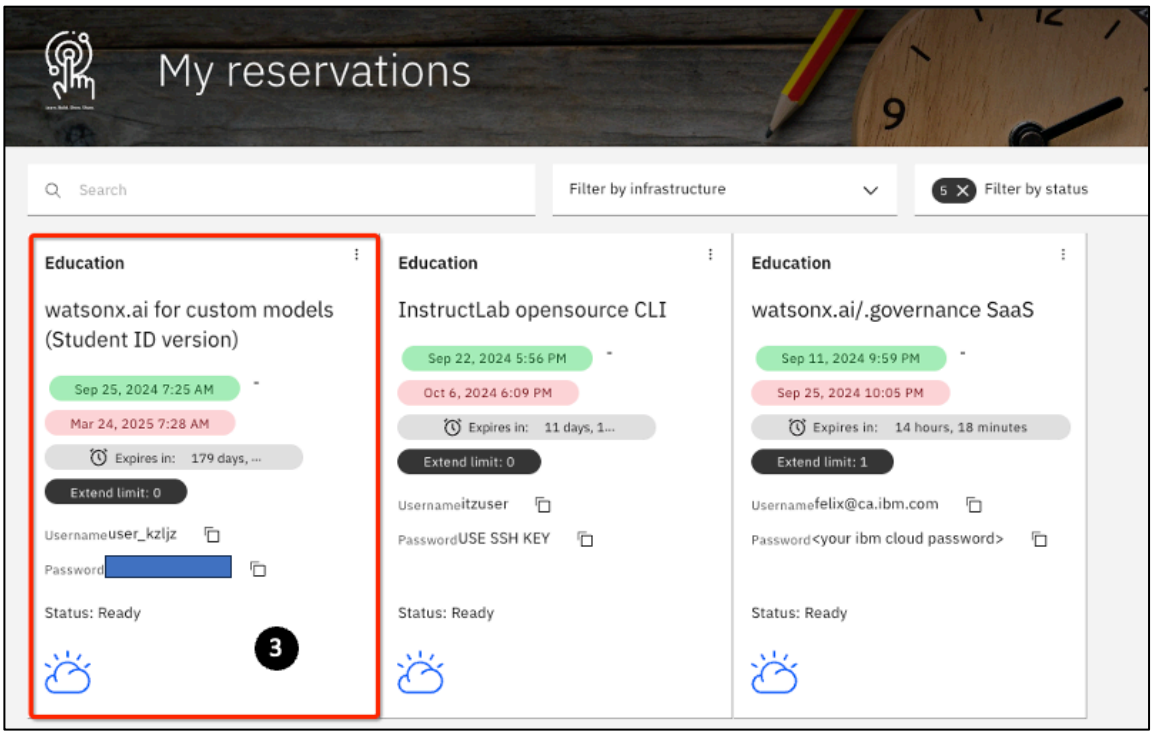


3. The **My reservations** page opens. Look for the **watsonx.ai for custom models** tile (the **Password** is intentionally redacted in the image below).

   You can copy the **Username** and **Password** and keep them handy, or you can always return to this page to get it.

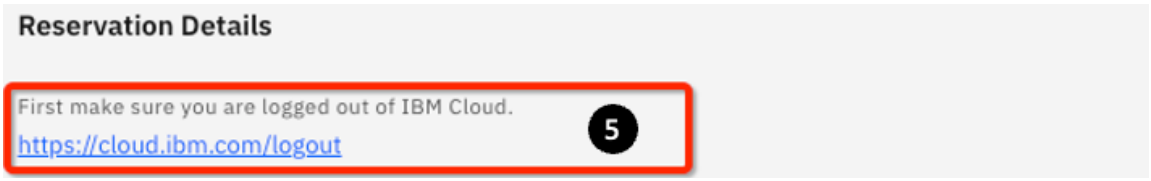
   In this example, the **Username** is **user_kzijc**.

© 2024 IBM Corporation

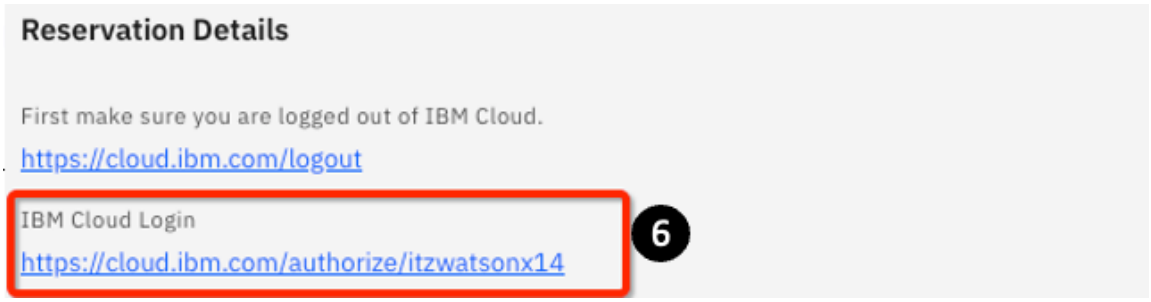


4. Click on the tile to open its details.

5. Scroll down to the **Reservation Details** section. Click https://cloud.ibm.com/logout to log out of other sessions.



6. Return to your browser tab with the reservation details. Click on the link provided for **IBM Cloud Login** (your link will be different from the example shown below).



7. A new login panel appears. Use the **Username** and **Password** provided in your TechZone reservation to log in.

© 2024 IBM Corporation

8. Click **Sign in**

9. You have now logged into the IBM Cloud using the Student ID. You can stay logged in, or simply log back in later using the same procedure.

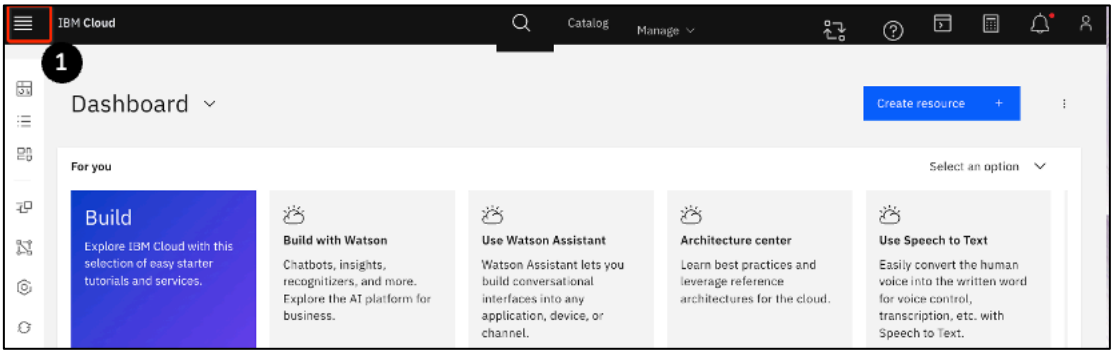
   You must always be logged into your TechZone reservation in this manner for this lab.

5.4.3 Creating a project in watsonx.ai

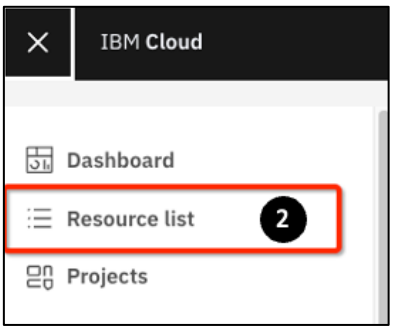
Now you will create a sandbox project that will include a Cloud Object Storage (COS) instance that you can use to upload your model.

If you have logged out of IBM Cloud (or it has timed out), you can log back in using the process in Section 5.4.2. Do NOT login with your IBM ID.

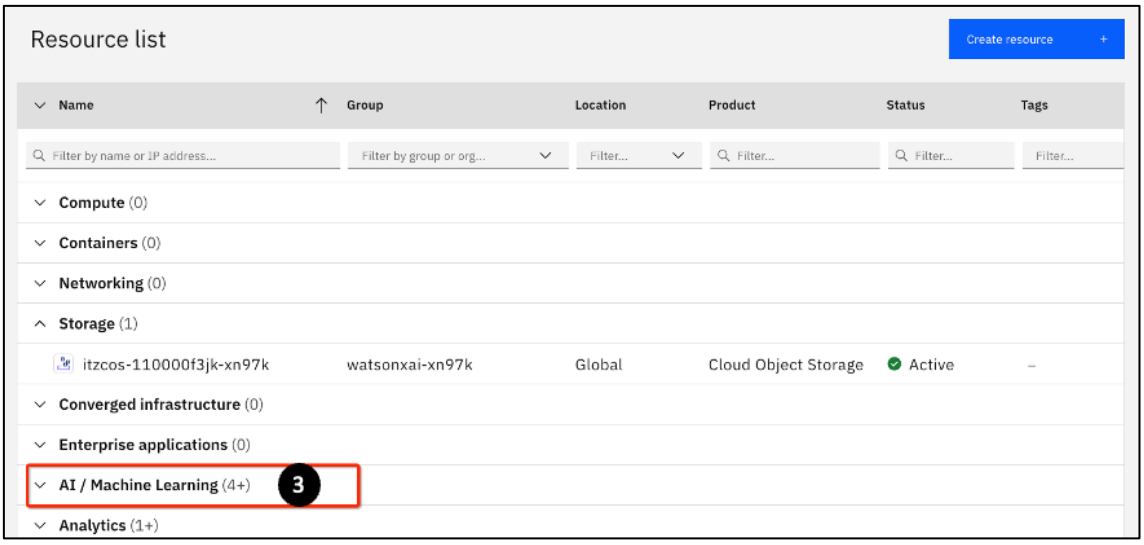
1. On the **Dashboard**, click on the main menu on the upper left.



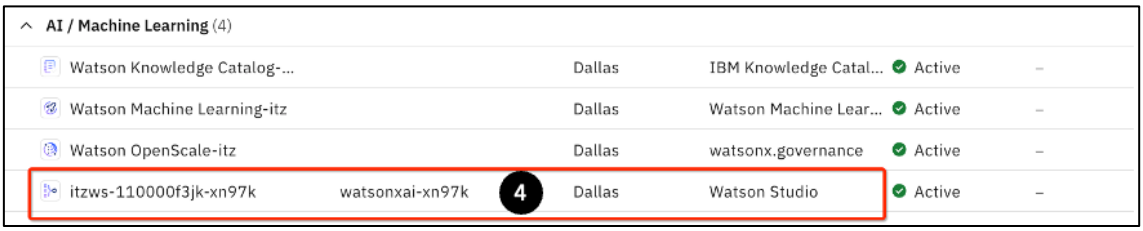
2. Click **Resource list**



3. Expand the **AI/Machine Learning** section by clicking on it.



4. Click on the **Watson Studio** instance highlighted below (your entry will have a different itzws-*** label).



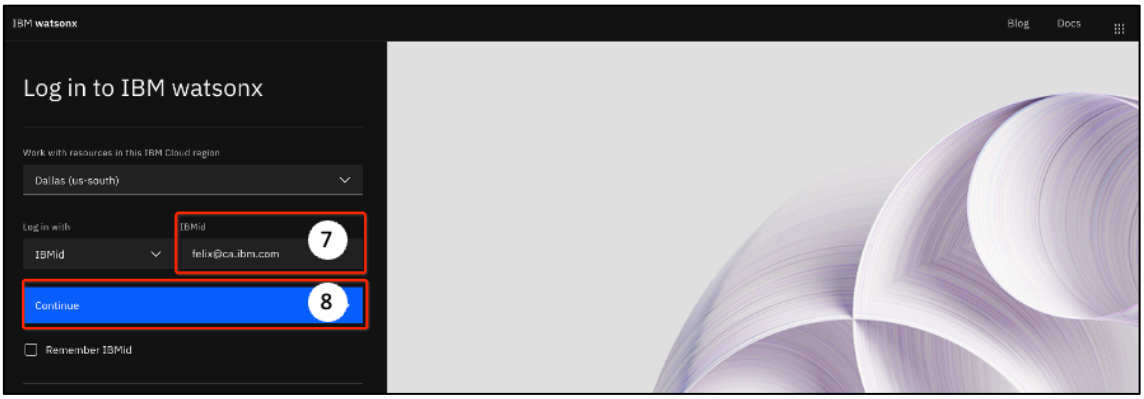
5. On the **Watson Studio** page, click to expand and see options for **Launch in**.

© 2024 IBM Corporation

6. You will see 2 options: **IBM Clock Pak for Data** and **IBM watsonx.** Select **IBM watsonx**.

7. You may be asked to log into the IBM cloud. If so, provide your IBMid this time only. Note that you have already logged into your Student ID (**user_kzijc**). This is an additional step to get you into the watsonx.ai console.

8. Click **Continue**.

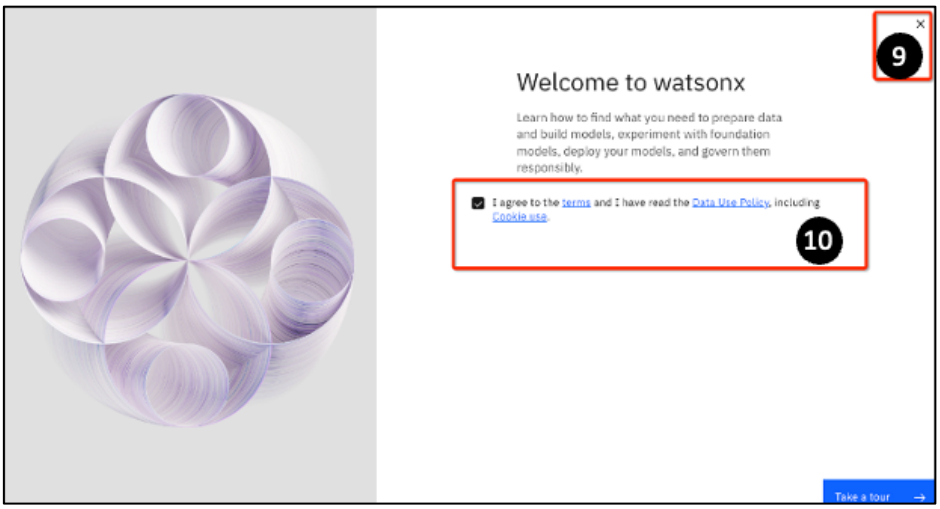


9. The **Welcome to watsonx** page may open, asking you to agree to terms. If it does, agree to its terms.

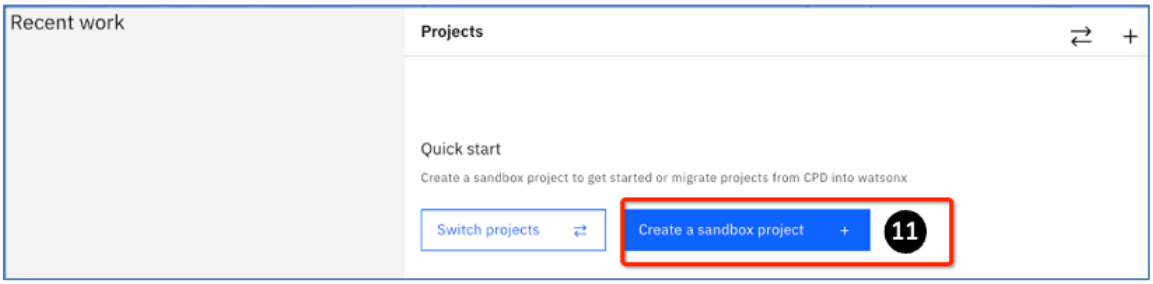
   If TechZone assigns an account you have used before, your **Welcome to watsonx** page may not ask you to agree to the terms.

10. In either case, click **X** to close this window.

© 2024 IBM Corporation



11. You are now in the watsonx.ai console. Scroll down to the **Projects** area and click **Create a sandbox project**.



12. A project called **notset's sandbox** is created. Note that you are using a Student ID, not your personal ID so the project would not have your name (or an **itz-watsonx** prefix like in other watsonx.ai L3 labs.

    The sandbox project may take some time to create, so please be patient.

When you create the sandbox project, a Cloud Object Storage (COS) bucket is created. You will upload the model files into this bucket in the next section.

© 2024 IBM Corporation

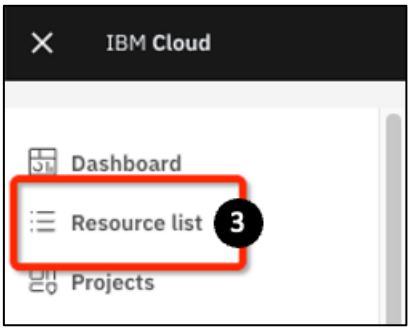
5.5 Uploading the model to Cloud Object Storage

Now you will upload the model files to the COS bucket associated with your project.

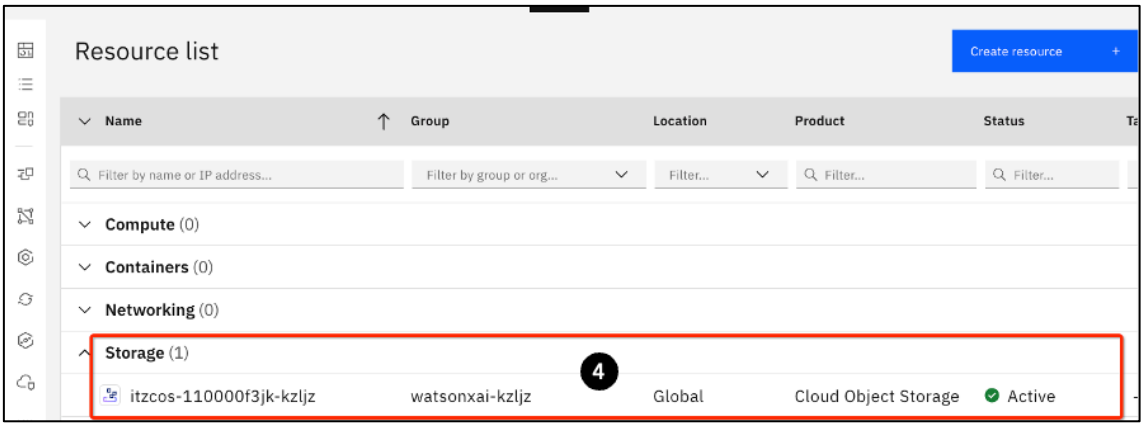
1. If you have logged out of IBM Cloud (or it has timed out), you log back in using the process in Section 5.4.2. Do NOT login with your IBM ID.

2. On the **Dashboard**, click on the main menu on the upper left.

3. Click **Resource list** on the slide-out.



4. Expand the **Storage** section by clicking on it.

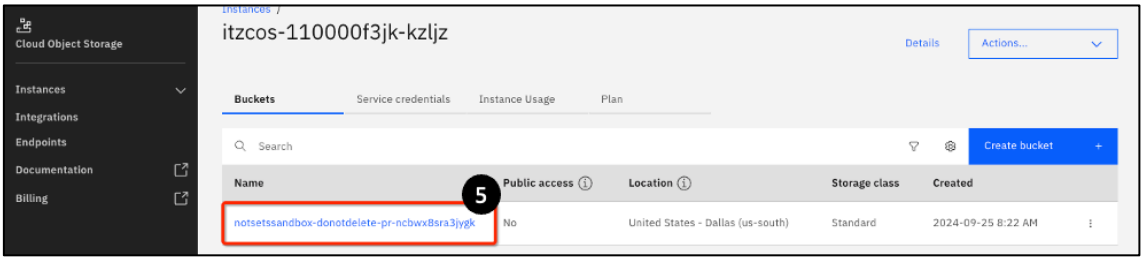


The entry with **Name** similar to **itzcos-*-*** (as shown above) is the bucket available to you for this lab. In this example, it is **itzcos-110000f3jk-kzliz.** Click on this **itzcos-*-*** instance.

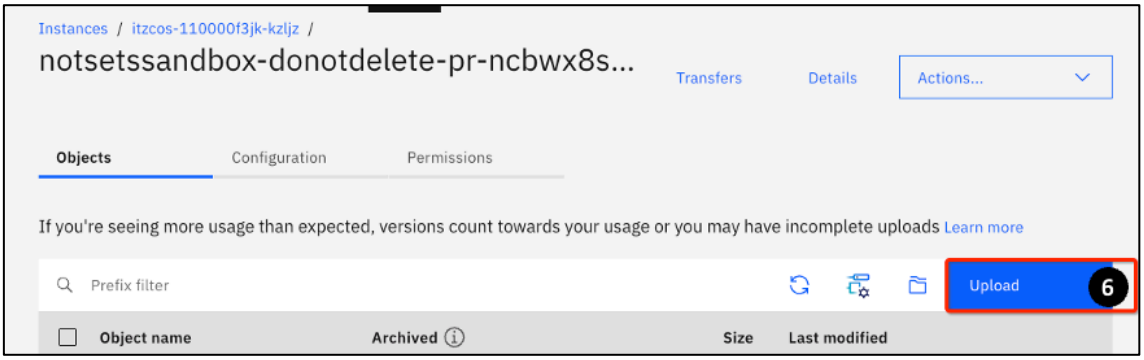
© 2024 IBM Corporation

5. The **itzcos-*-*** instance page opens, showing the bucket associated with it. There might be multiple buckets, but you should see an entry with a name like **notsetsandbox-donot-delete-***

   Click on that bucket to open it.



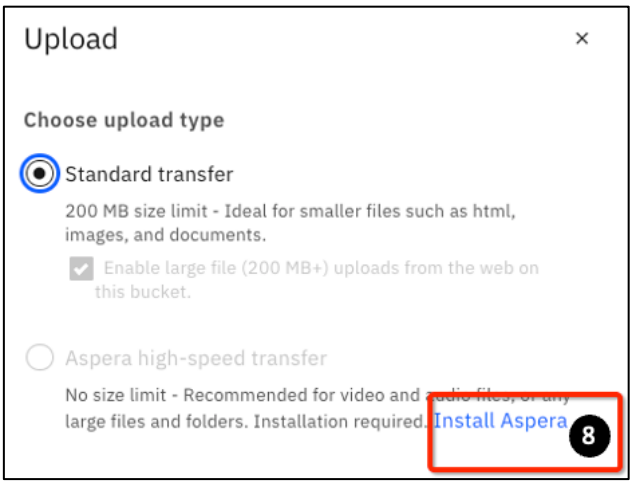
6. Click **Upload** on the right.



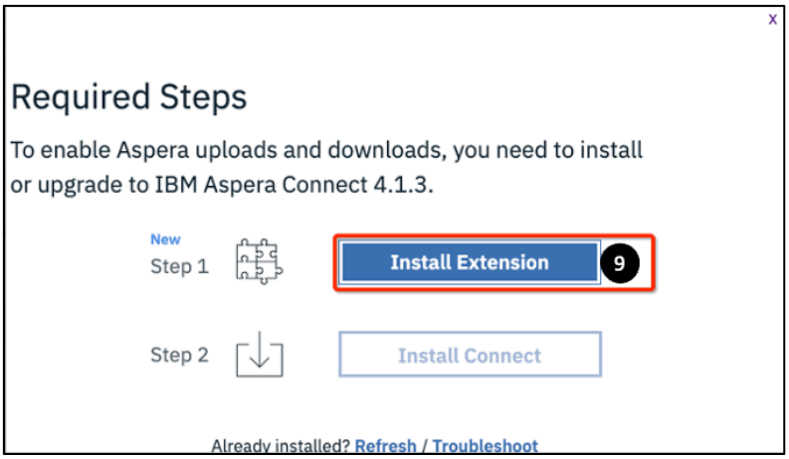
7. The **Upload** slide-out opens. Notice how the **Standard transfer** is proposed for file transfers within the 200 MB limit. For this lab, since the **flan-t5-base** model is bigger you should use the **Aspera high-speed transfer** option. If you can select this option, select it and skip to Step 21

8. Click the **Install Aspera** hotlink associated with the **Aspera high-speed transfer** option.

© 2024 IBM Corporation



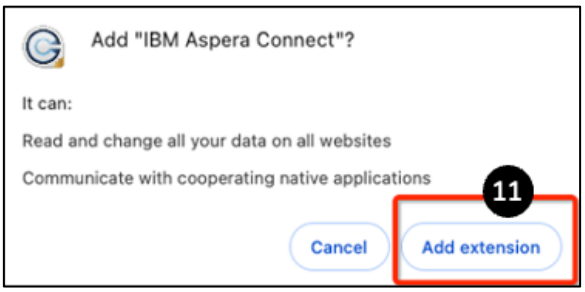
9. The **Required Steps** page appears. Click **Install Extension**. If the Extension is already installed, the **Install Connect** button is highlighted instead. If so, skip to Step 13.



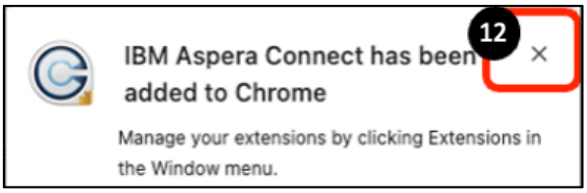
10. The **Chrome web store** page opens. Click **Add to Chrome**.

11. On the **Add "IBM Aspera Connect"** dialog box, Click **Add extension.**

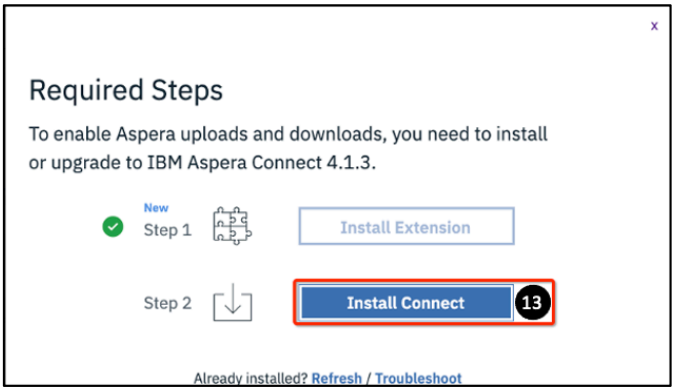
© 2024 IBM Corporation



12. You will see a message that Aspera has been successfully installed. Click **x** to dismiss this window.



13. You are back to the **Required Steps** page. Click **Install Connect**.



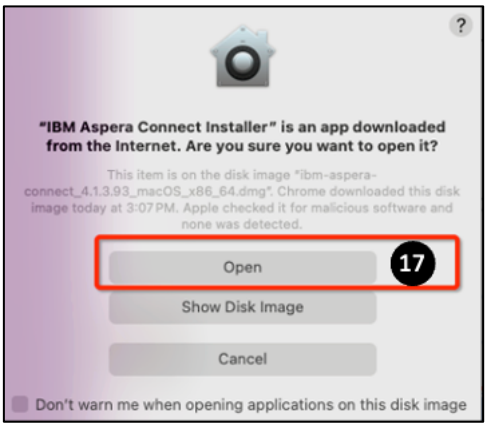
14. You will see the **IBM Aspera Connect Installer** downloaded. Depending on the OS, you may see different files. In the case of Mac OS, you see a file like **ibm-aspera-connect_4.1.3.93_macOS_x86_64.dmg** (version number 4.1.3.93 may differ).

15. Navigate to the **Downloads** (or your equivalent) folder and click on the file.

16. A folder opens up. Open the **IBM Aspera Connect Installer** icon.

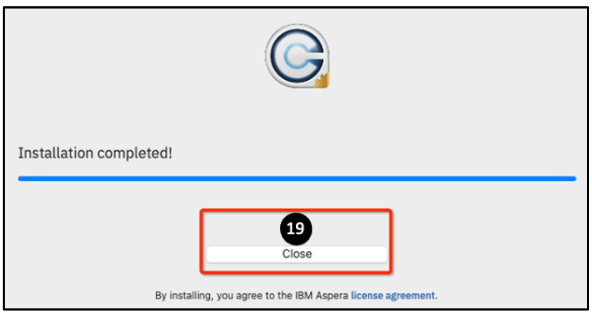
© 2024 IBM Corporation

17. On Mac OS, you will see a warning panel. Click **Open**.



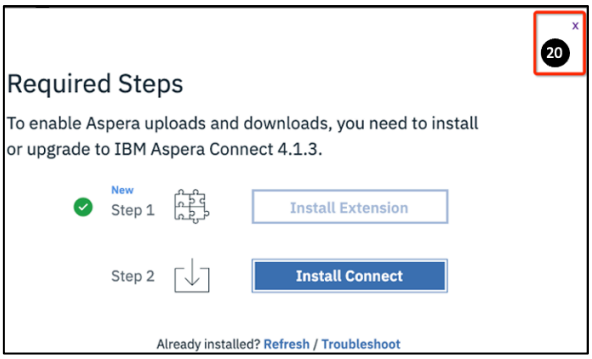
18. The **To install IBM Aspera Connect** page opens. Click **Start**.

19. When installation is completed, click **Close**.



© 2024 IBM Corporation

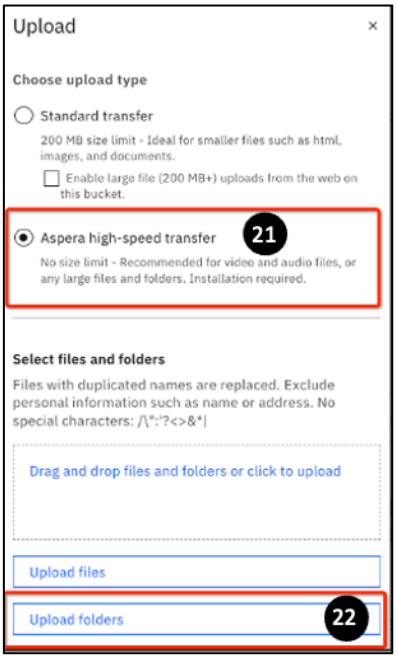
20. Click **X** to close the **Required Steps** window (if still open).



21. On the watsonx.ai console, click **Upload** again. Now, the **Aspera high-speed transfer** option is available to you. Select **Aspera high-speed-transfer**.

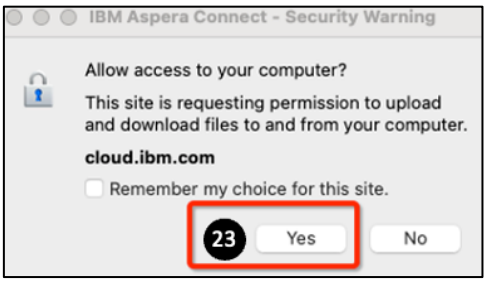
    Note: if you have successfully installed the plug-in and the Aspera image but the **Aspera high-speed transfer** option is still not available, the plug-in may be blocked by the brower.

22. Click **Upload folders**.



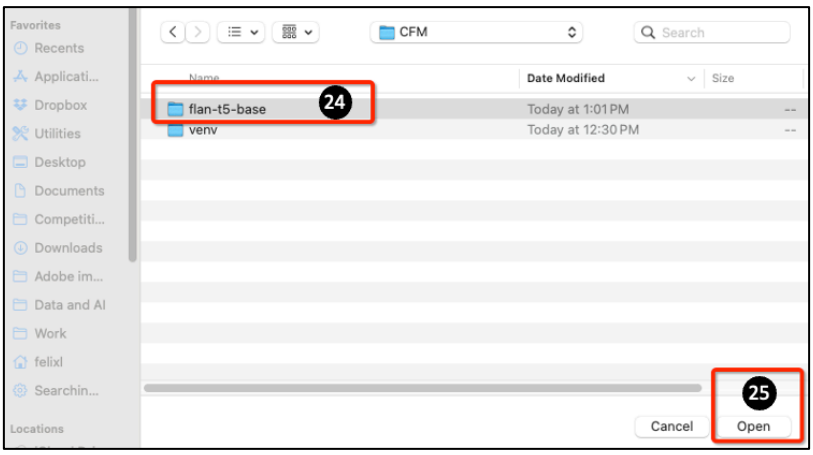
23. You will be asked if you **Allow access to your computer**. Click **Yes**.

© 2024 IBM Corporation



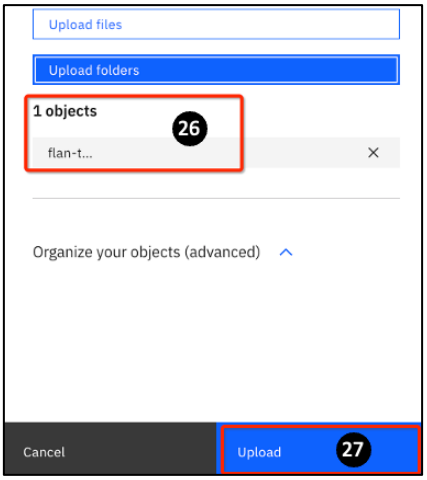
24. Navigate to the **CFM** folder (**/Users/felixl/Downloads/CFM** in this example) and select the **flan-t5-base** folder.

25. Click **Open** to upload.



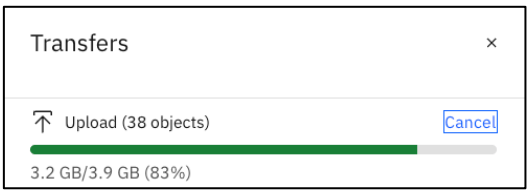
26. You will see "**1 objects**" as shown below - the folder **flan-t5-base**.

27. Click **Upload**.

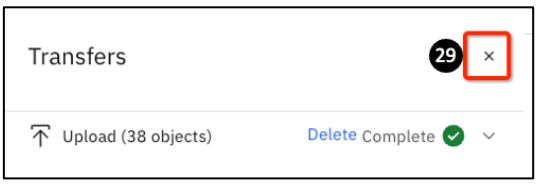


28. You will see the following message. Aspera provides high-speed transfer, so this should be quick. It does depend on your connection speed to IBM Cloud.

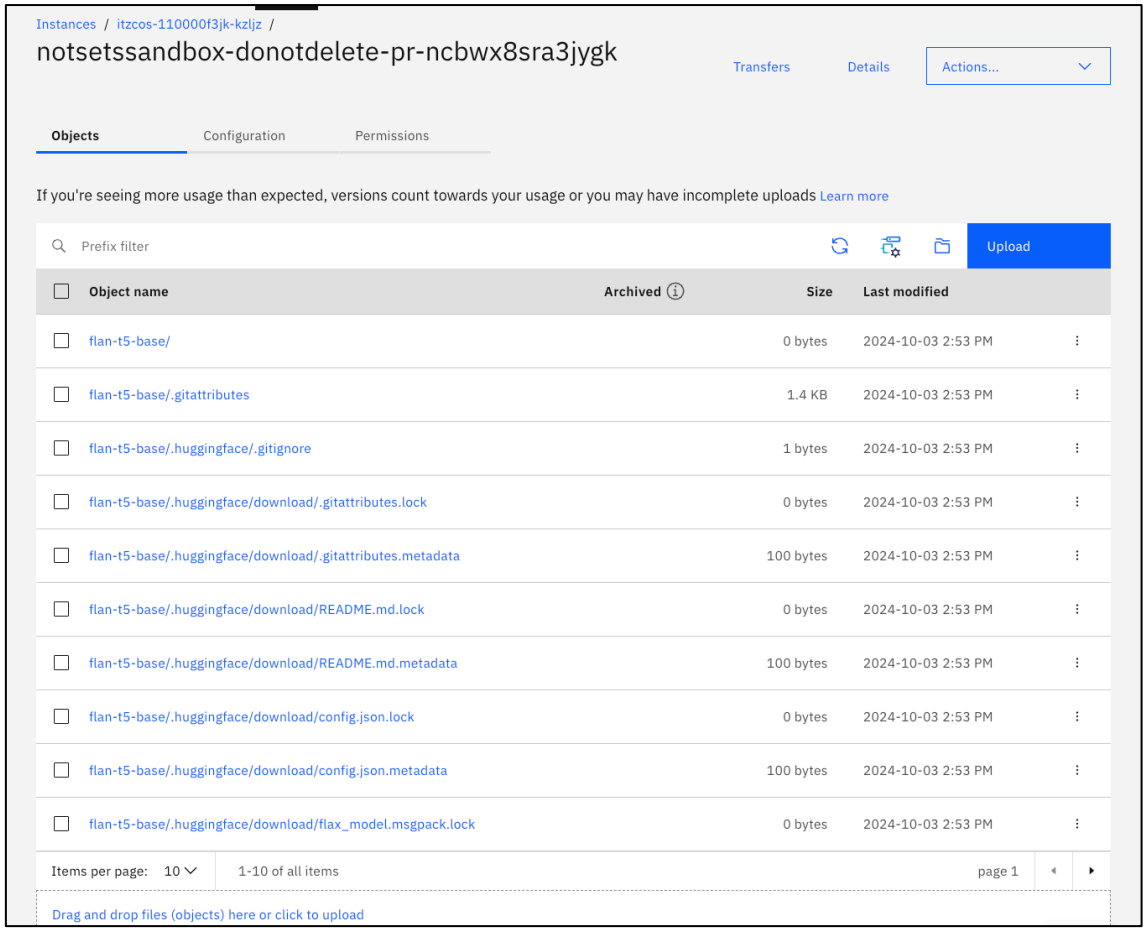
© 2024 IBM Corporation



29. When the upload is completed, you will see the image below. Click **X** to close the **Transfers** panel.



30. You will see a lot of files now in your COS bucket. You will have multiple pages.

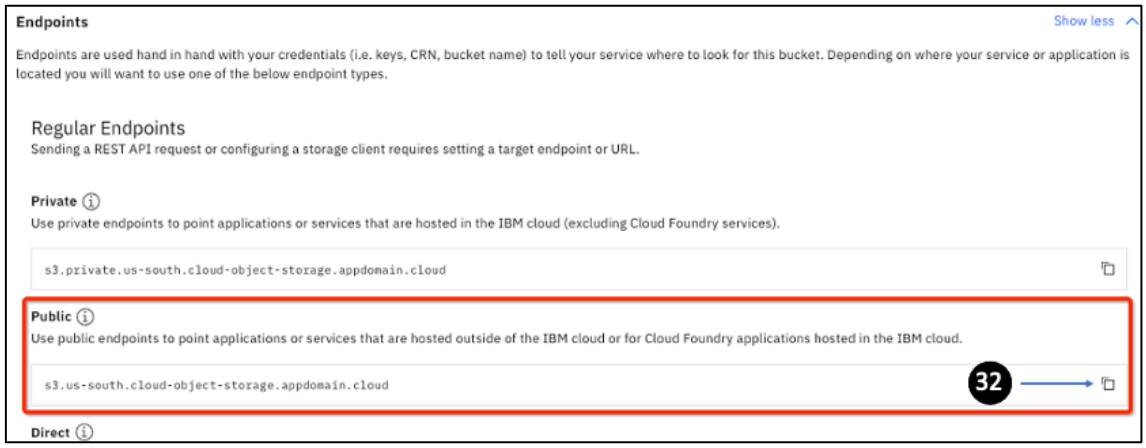


You are now ready to connect to this from your watsonx.ai instance. To do so, you need additional information on the COS bucket.

© 2024 IBM Corporation

31. Click the **Configuration** tab.

32. Scroll down to the **Endpoints** section and find the **Public** endpoint to this bucket. Click the **Copy** button on the right of the entry to put the information on the clipboard. Save this information somewhere you can retrieve later.



In this example the public endpoint is
s3.us-south.cloud-object-storage.appdomain.cloud

© 2024 IBM Corporation

5.6 Creating a connection to COS from watsonx.ai

You have uploaded a model into the COS instance associated with your TechZone reservation. Now you need to create a connection from watsonx.ai to access the model's files.

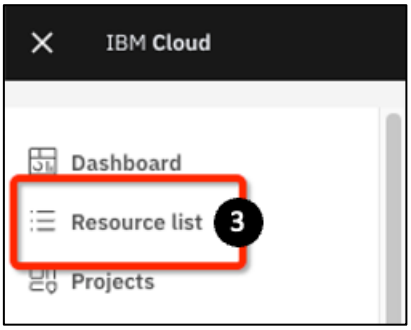
5.6.1 Getting your Access key and Secret key

When you connect to your COS instance to access your model files for Custom Foundation Model, you need to connect using **Access key and Secret key.**

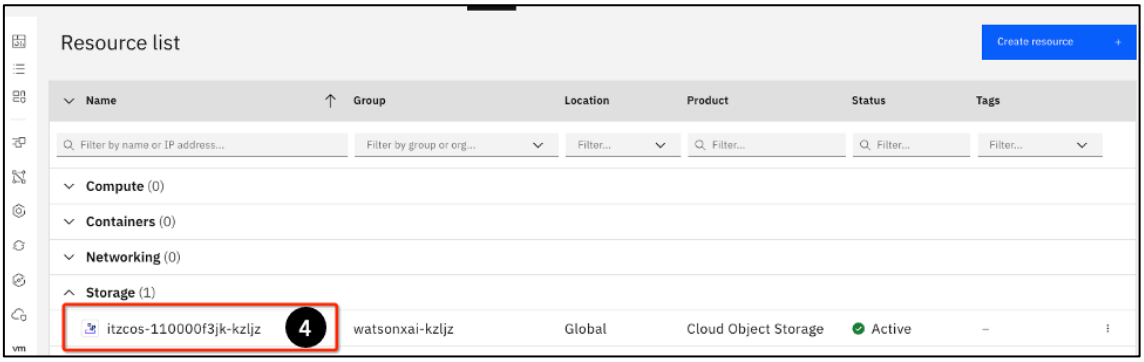
1. If you have logged out of IBM Cloud (or it has timed out), you can log back in using the process in Section 5.4.2. Do NOT login with your IBM ID.

2. On the **Dashboard**, click on the main menu on the upper left.

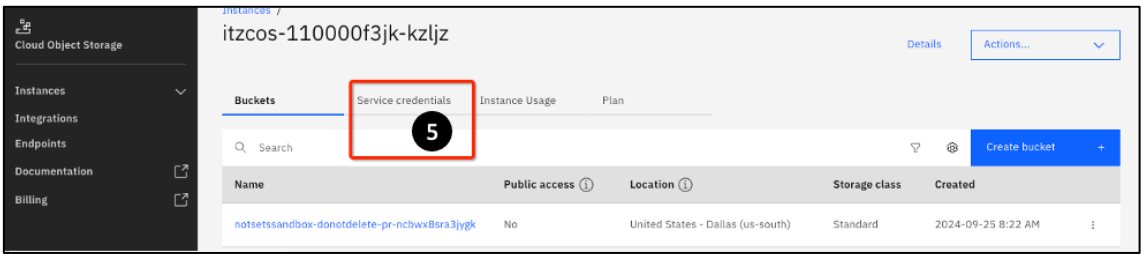
3. Click **Resource list** on the slide-out



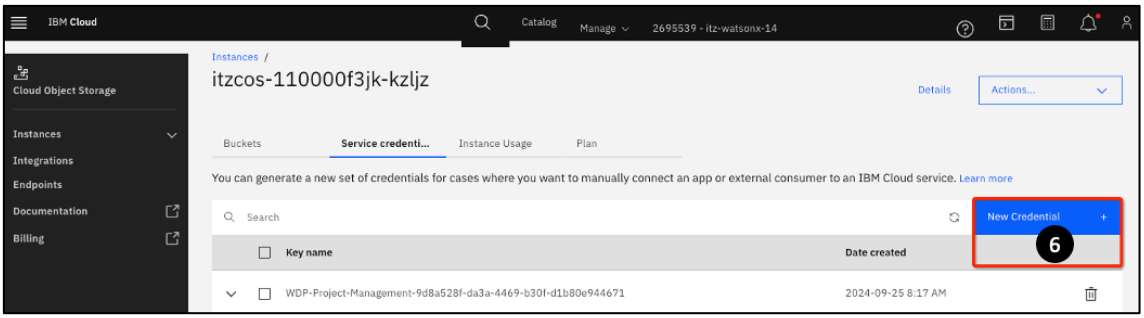
4. Expand the **Storage** section and click on the COS instance with name like **itzcos-***.



5. Select the **Service credential** tab (if not already selected)



6. You may already have several keys. Click **New Credential** to create a new set.



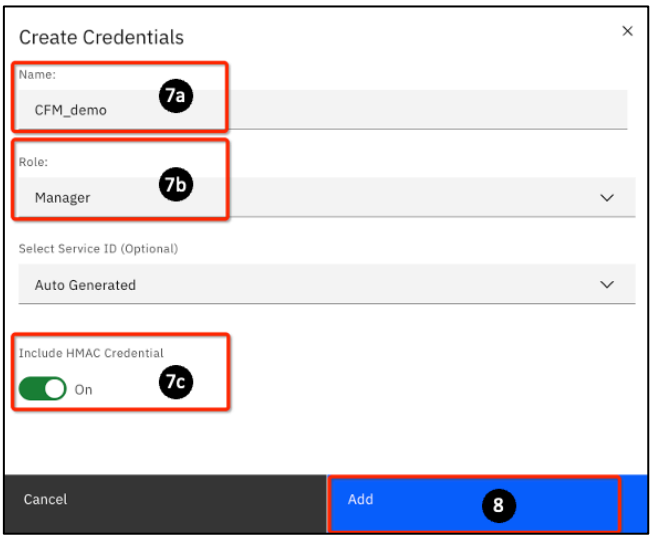
7. The **Create Credentials** page opens. Use the following values:

    a. For **Name**: use **CFM_demo**

    b. For **Role**: select **Manager**

    c. For **Include HMAC Credential**: set it **On** by clicking on the button and slide it to **On**.

8. Click **Add**.



© 2024 IBM Corporation

9. Scroll down (if necessary) to find the **CFM_demo** key and click it to get its details.

10. Click the **copy** icon to place that information on your clipboard. Paste it in a file so you can easily retrieve later.

11. The content of this sample key is pasted below. Note the values of **access_key_id** and **secret_access_key** (you need the values within the quotation marks). In this example:

    a. **access_key_id** is 0912c828a1704903884590e9eb46f1d3
    b. **secret_access_key** is 11e50e250e0440ba40b46add3cfbc6e8146dfb5f04e0570e


```
{
  "apikey": "yb1y0C5P07m_dtF0aPe-L3oz1pCPyO8nyQ4k7TC1op_0",
  "cos_hmac_keys": {
    "access_key_id": "0912c828a1704903884590e9eb46f1d3",
    "secret_access_key":
"11e50e250e0440ba40b46add3cfbc6e8146dfb5f04e0570e"
  },
  "endpoints": "https://control.cloud-object-storage.cloud.ibm.com/v2/endpoints",
  "iam_apikey_description": "Auto-generated for key crn:v1:bluemix:public:cloud-
object-storage:global:a/baf3554ebf3c4225ac4c4e9efb79516b:9d8a528f-da3a-4469-
b30f-d1b80e944671:resource-key:0912c828-a170-4903-8845-90e9eb46f1d3",
  "iam_apikey_id": "ApiKey-b6c65d0c-fbca-4939-802b-b1e001bf4f32",
  "iam_apikey_name": "CFM_demo",
  "iam_role_crn": "crn:v1:bluemix:public:iam::::serviceRole:Manager",
  "iam_serviceid_crn": "crn:v1:bluemix:public:iam-
identity::a/baf3554ebf3c4225ac4c4e9efb79516b::serviceid:ServiceId-a9e0b637-
f363-4193-82b6-9f4310216b61",
  "resource_instance_id": "crn:v1:bluemix:public:cloud-object-
storage:global:a/baf3554ebf3c4225ac4c4e9efb79516b:9d8a528f-da3a-4469-b30f-
d1b80e944671::"
}
```


You will use these values when you create the connection to the COS instance.

© 2024 IBM Corporation

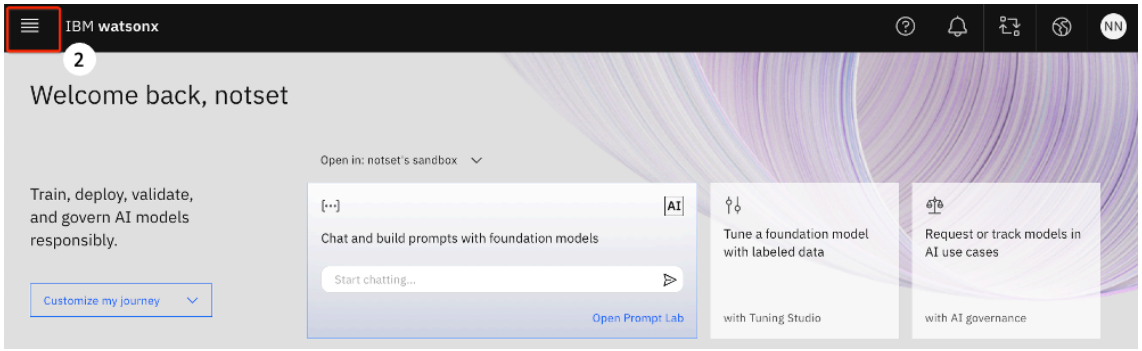
5.6.2 Creating a Deployment Space and connection to COS using Access key and Secret key

You can deploy a model to a project or to a deployment space. A deployment space is not associated with a project. Assets from different projects can be deployed in the same space. In this section, you will

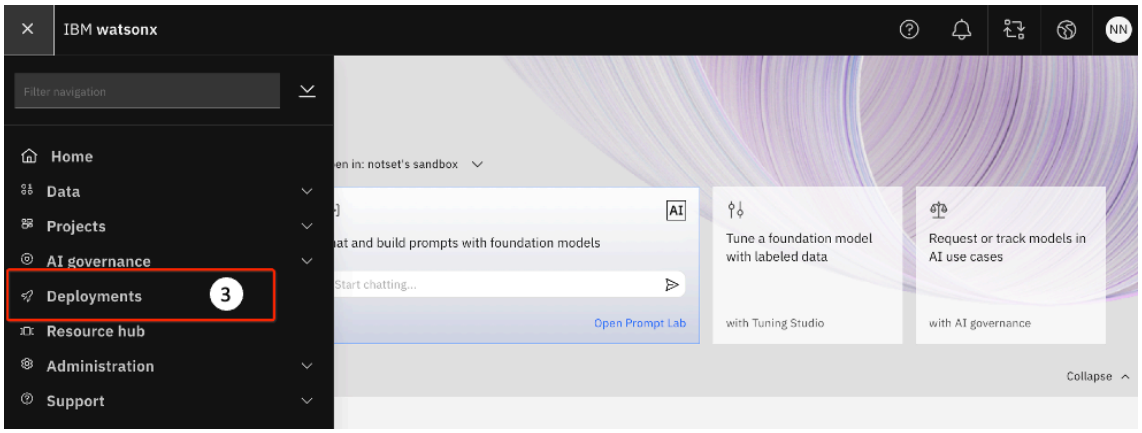
- Create a deployment space
- Define a connection from your deployment space to your COS instance.

1. You can log back into watsonx.ai using Steps 1-8 of Section 5.4.3.

2. The watsonx.ai main console opens. Click the main menu on the upper left.



3. On the slide-out, click **Deployments**.



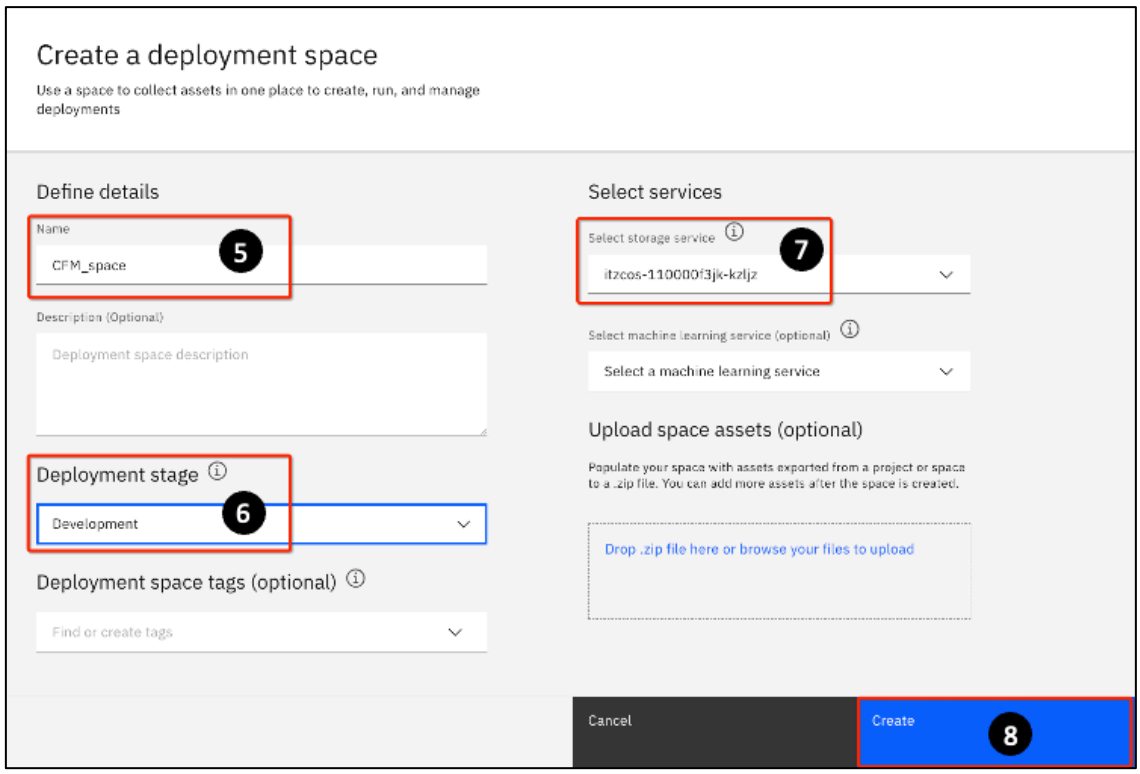
4. The **Deployments** page opens. Click the **New deployment space** button.

5. The **Create a deployment space** page opens. Enter **CFM_space** as **Name**.

6. Select **Development** for **Deployment stage**.

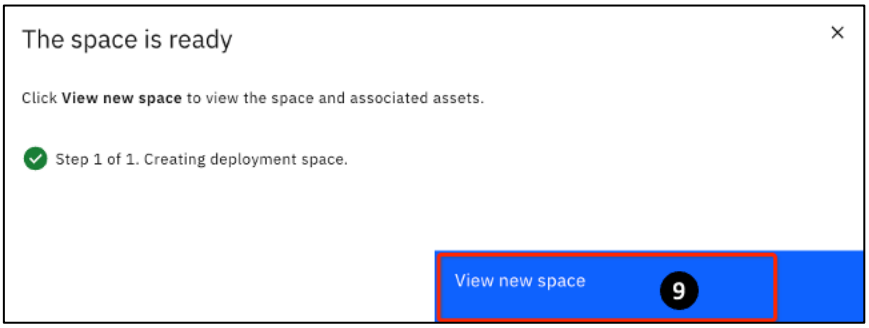
7. The **Select storage service** field should be automatically populated if you only have one COS instance. If you have more than one, select the one identified in Step 4 of Section 5.5 (with a name like **itzcoz-***).

8. Leave everything else as is. Click **Create**.

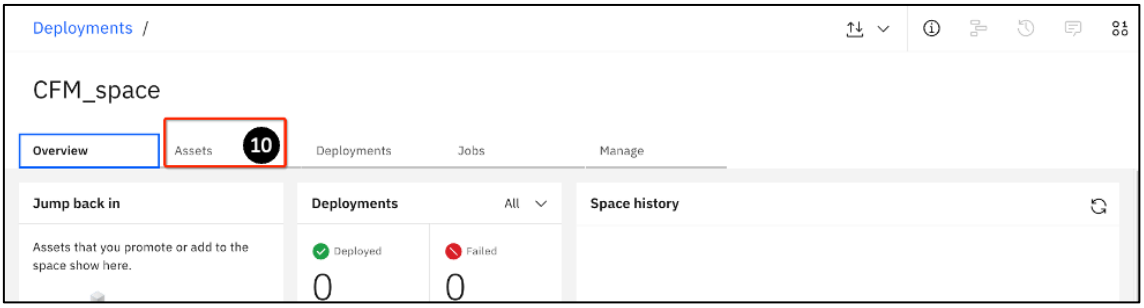


9. It takes some time to create a depolyment space. You will see this:

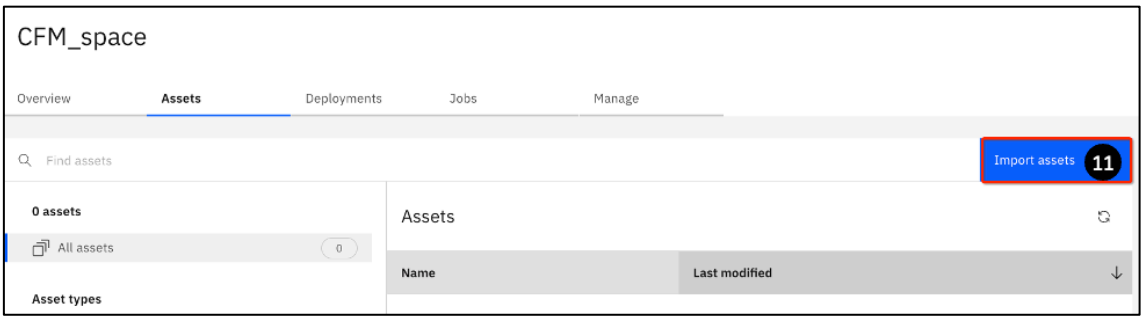
When the process is completed, the **View new space** button becomes active. Click **View new space**.



10. The **CFM_space** page opens. Click the **Assets** tab.

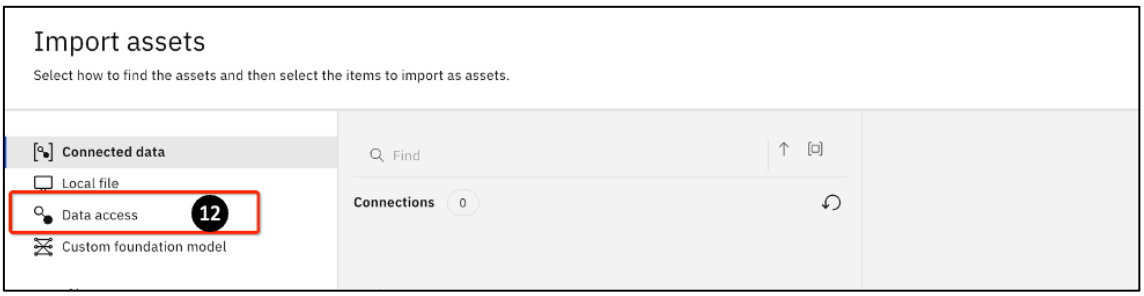


11. On the **Assets** tab, click **Import assets**.



12. The **Import assets** page opens. The first thing you need to do is to define data access to your COS instance where the model files have been uploaded. Click **Data access** on the left.

© 2024 IBM Corporation



13. Click the **Connection** tile.

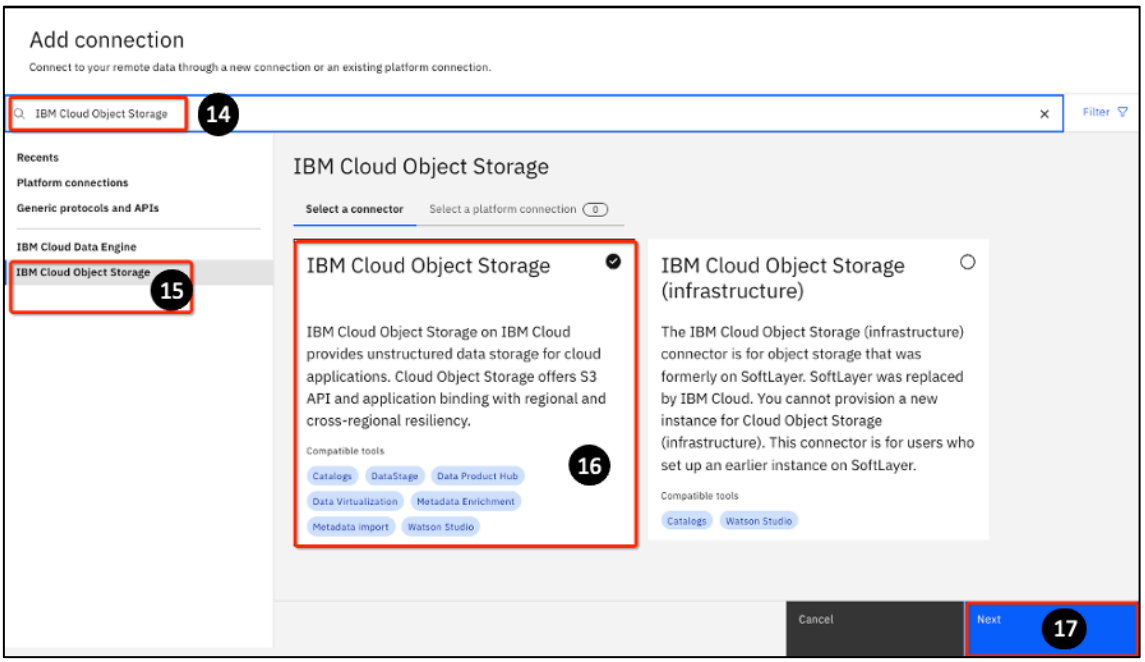
14. The **Add connection** page opens. If you do not see **IBM Cloud Object Storage**, go to the **Search** field at the top left and enter **IBM Cloud Object Storage.** The list will dynamically filter.

15. Click on **IBM Cloud Object Storage**.

16. You may see 2 tiles: **IBM Cloud Object Storage** and **IBM Cloud Object Storage (infrastructure)**. Select **IBM Cloud Object Storage** (if not already selected).

17. Click **Next**.

© 2024 IBM Corporation



18. On the **Connect to a data source** page:

    a. DO NOT click **Select instance**. If you do, the connection will fail.

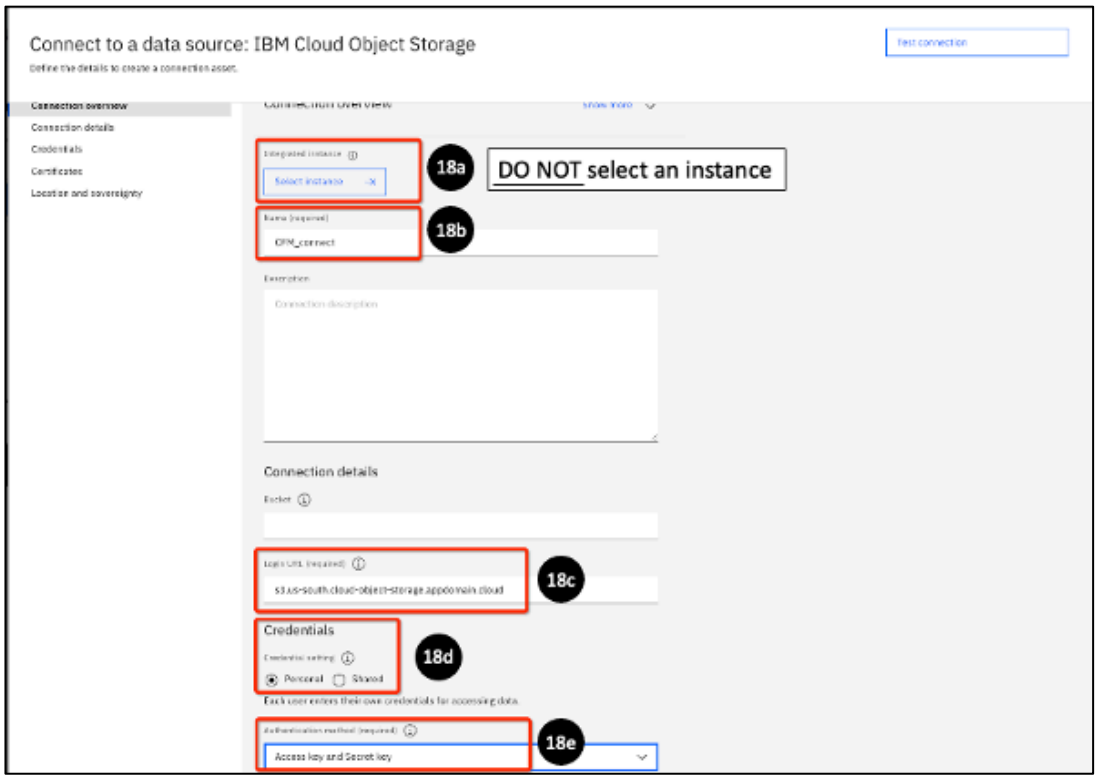
    b. For **Name**: enter **CFM_connect**

    c. For **Login URL**: enter the value you obtained in Step 32 of Section 5.5. For example: **s3.us-south.cloud-object-storage.appdomain.cloud**

    d. For **Credentials**, click **Personal.**

    e. For the **Authentication method**, select **Access key and Secret key.**

© 2024 IBM Corporation

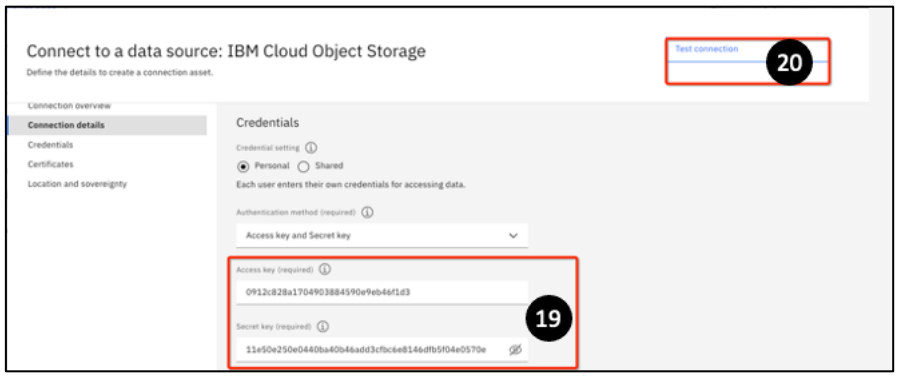


19. Two additional fields will appear when you select **Access key and Secret key**. Enter the values you received for your **CFM_demo** in Step 11 of Section 5.6.1. In this example:

    Access key: 0912c828a1704903884590e9eb46f1d3

    Secret key: 11e50e250e0440ba40b46add3cfbc6e8146dfb5f04e0570e

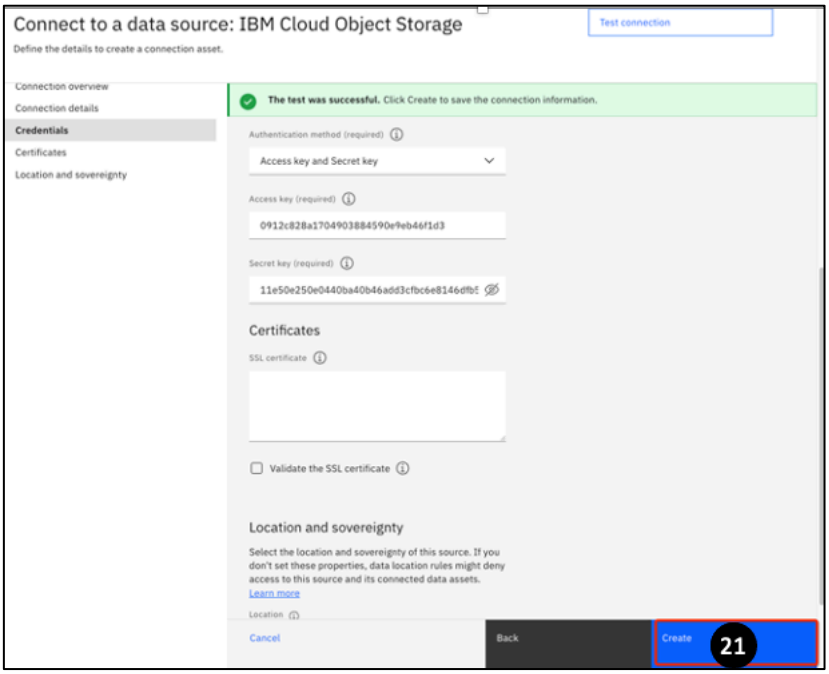
20. Click **Test Connection**.



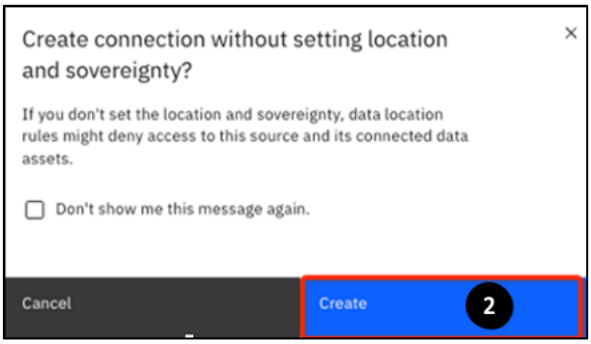
You should see a successful test result:

© 2024 IBM Corporation

21. Click **Create**.

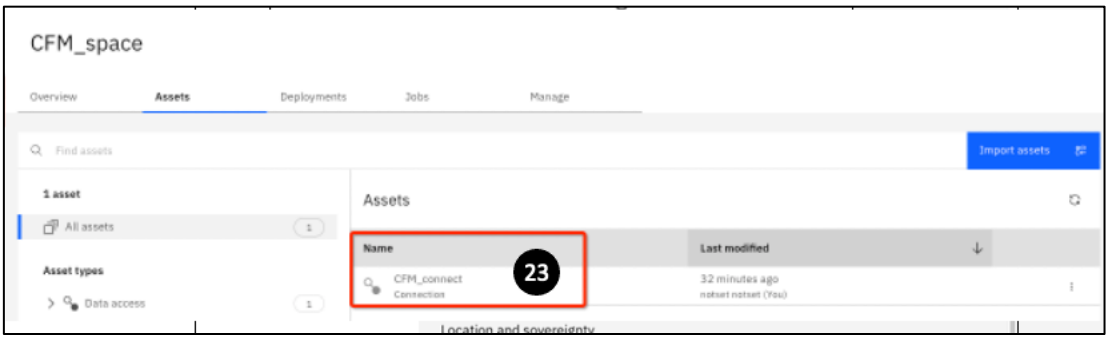


22. You will get a warning. You can ignore it and click **Create**.



23. On the **CFM_space**, the newly created **CFM_connect** asset appears.

© 2024 IBM Corporation



You have created a deployment space and defined a connection to the COS instance in this section.

© 2024 IBM Corporation

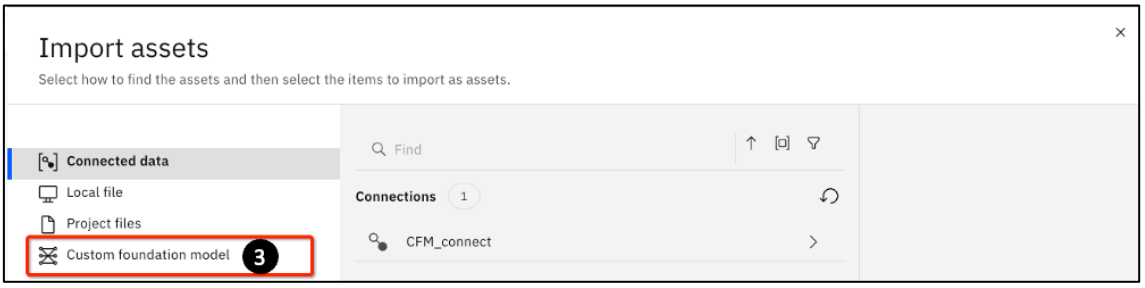
5.7 Creating a model asset

In this section, you import your model asset into the newly created deployment space.

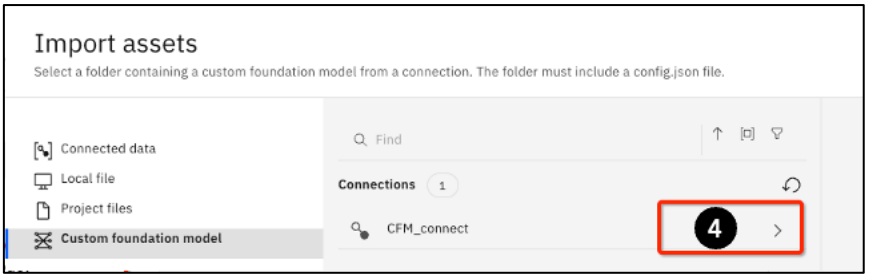
1. On the **CFM_space Assets** tab, click **Import assets**.

2. The **Import assets** page opens (probably to the default **Connected data** menu item). You should see the **CFM_connect** connection.

3. Click the **Custom foundation model** menu item on the left.



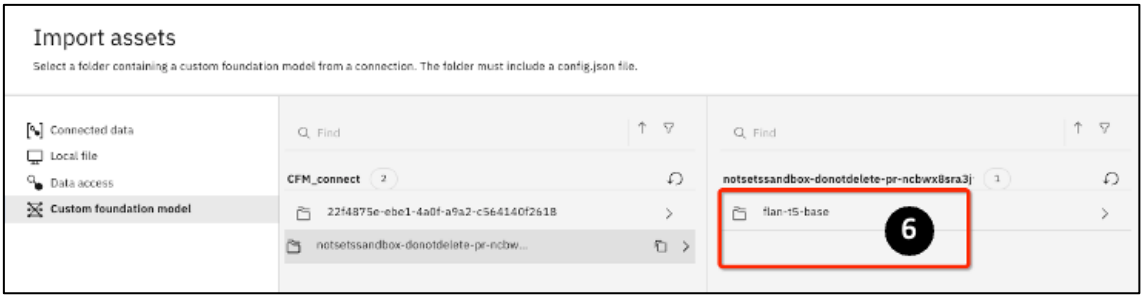
4. You will see the **CFM_connect** connection created in Step 18 of Section 5.2.6. Click on it to expand.



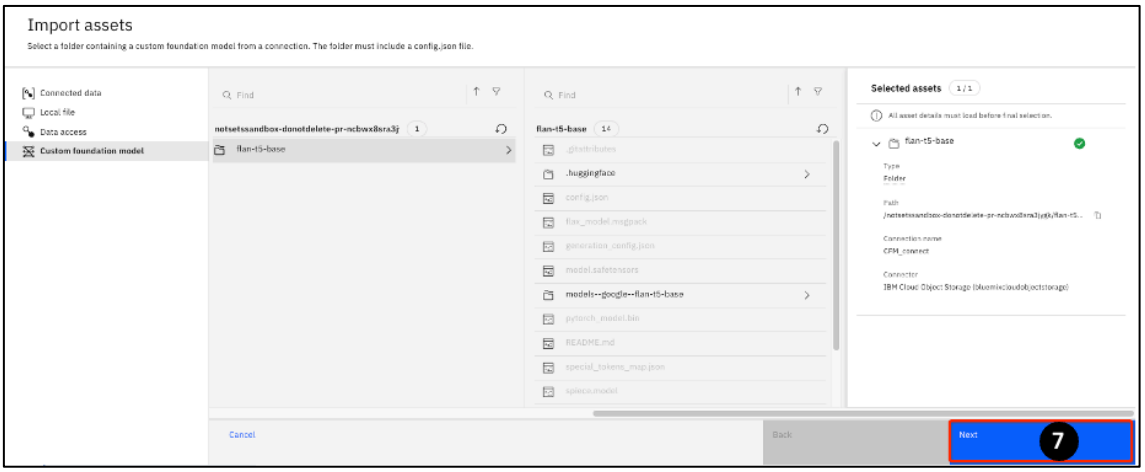
5. You should see the **CFM_connect** COS bucket. If you have more than one bucket, ensure you select the correct one (identified in Step 4 of Section 5.5). Click on it to expand.

© 2024 IBM Corporation

6. You may need to scroll right to see the details. You should see the content of the bucket: the folder **flan-t5-base**. Click on this folder.



7. The folder's content is shown. Click **Next** to import the entire folder.



8. The **Import a custom foundation model** page opens. Note that some values are filled in for you already:

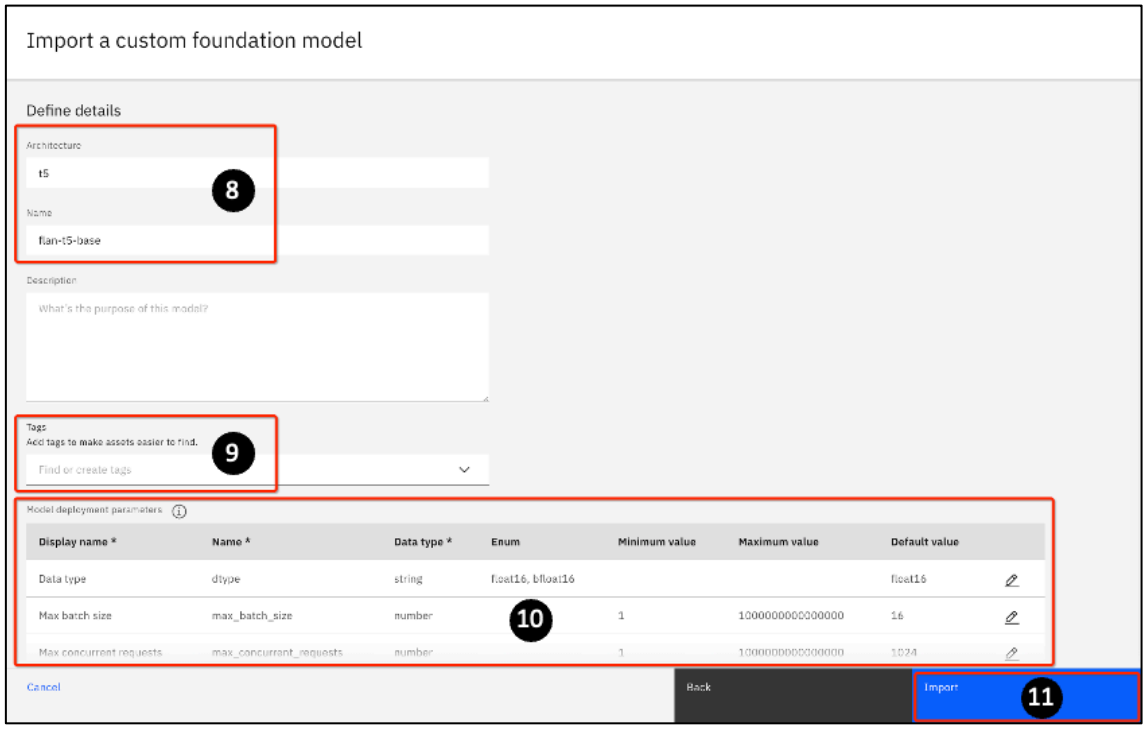
   Architecture: **t5**
   Name: **flan-t5-base** – leave it as is for the exercise. You may want to change the model name in a Proof of Experience (PoX) or other activities with a client to make it easier to remember.

9. Optionally, you can add tags – leave this empty for this exercise.

© 2024 IBM Corporation

10. Note the **Model deployment parameters** at the bottom of the page. You can update these by clicking on the associated pencil icon to the right. For this exercise, simply use the default values provided.

11. Click **Import**.



12. The model page for **flan-t5-base** opens. Note the **About this asset** information. You can provide changes/updates (such as adding a description) if you want. For now, leave this as is.

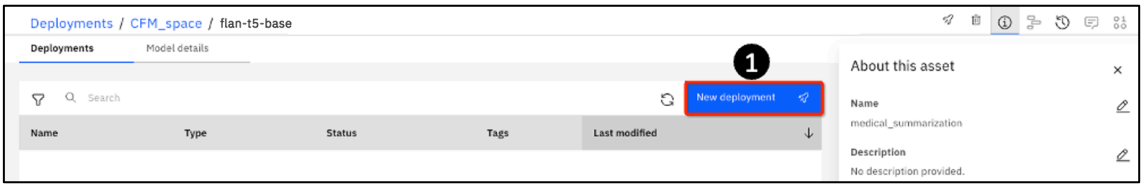
13. The model asset needs to be deployed before it is usable. You will do it in the next section. Do NOT close this window.

© 2024 IBM Corporation

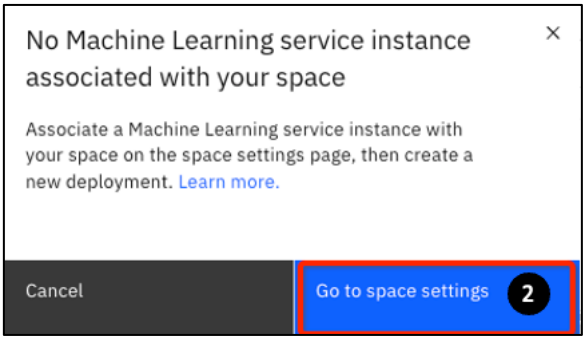
5.8 Deploying a model asset

Now, you will deploy the model that you have imported. This is required before the model can be used in watsonx.ai. Without this step, your model is basically just a set of files in a COS bucket. To deploy your model, you need to first create a user API key.

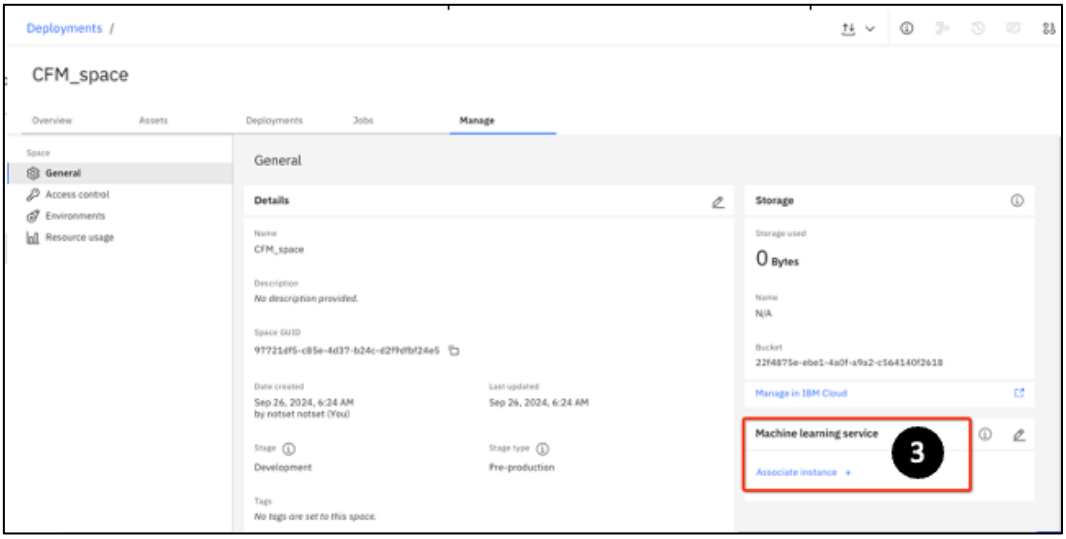
1. Click **New Deployment** on the **flan-t5-base** asset page.



2. You may get the following warning message about **Machine Learning Service instance**. If so, click **Go to space settings**. If not, skip to Step 10 below.



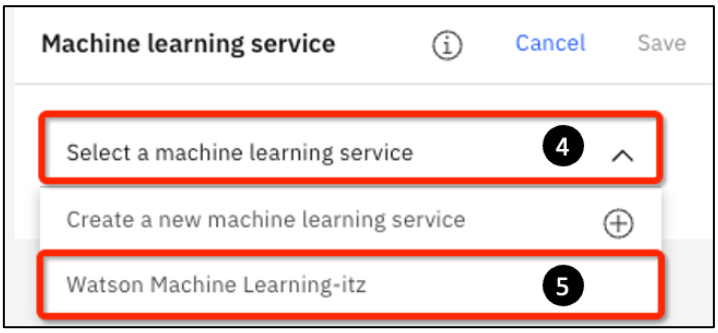
3. The **CFM_space** page opens. Click **Associate instance** from **Machine learning service** field on the right.



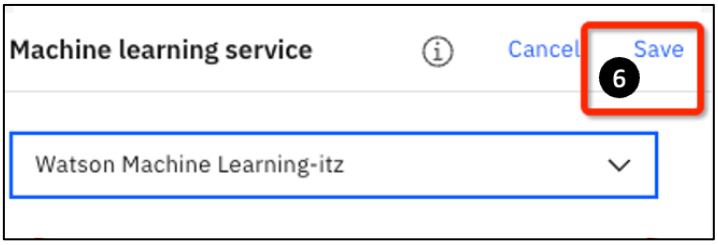
© 2024 IBM Corporation

4. Expand **Select a machine learning service** to show the list.

5. Select (you might have more than one) the **Watson Machine Learning-itz** service.

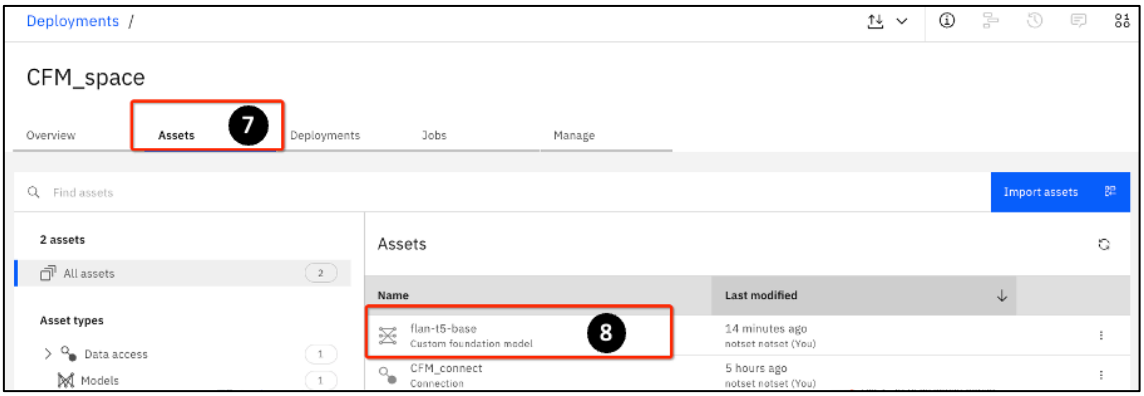


6. Click **Save**.



7. You are on the **Manage** tab. Click the **Assets** tab.

8. Click on the **flan-t5-base** asset.

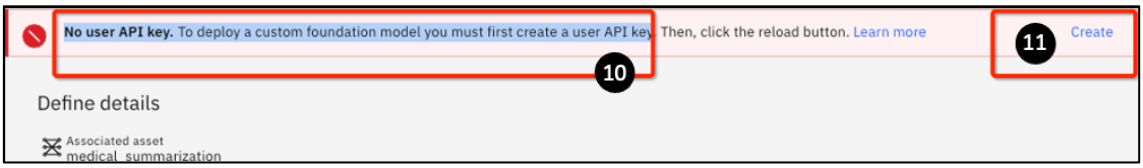


9. On the **flan-t5-base** page, click **New deployment**.

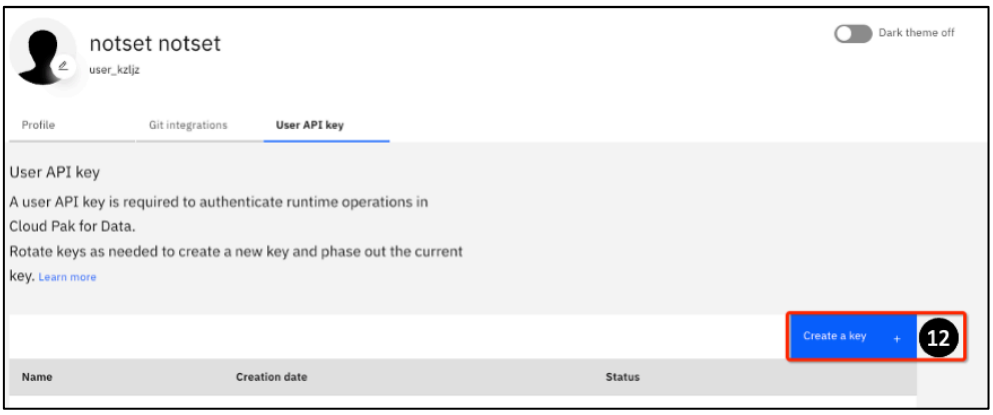
© 2024 IBM Corporation

10. The **Create a deployment** page opens. You may see a warning: **No user API key.** To deploy a custom foundation model, you must first create a user API key.

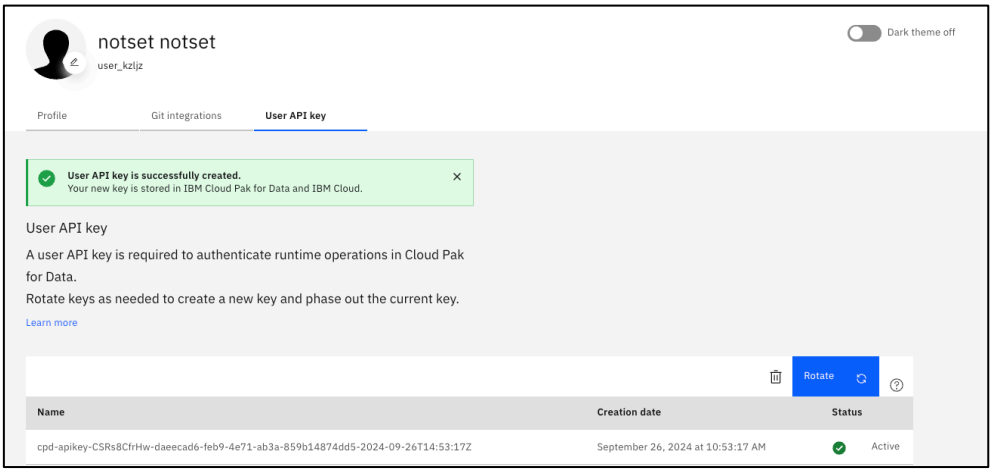
11. Click **Create**.



12. This will open the API creation page. Click **Create a key**. Return to the **Create a deployment** page (likely on a different tab of your browser).

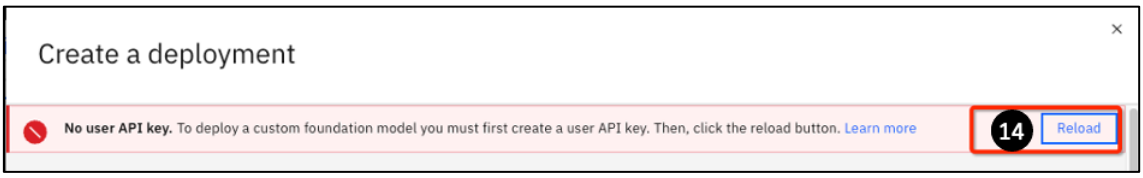


13. Your **User API key** will be created.

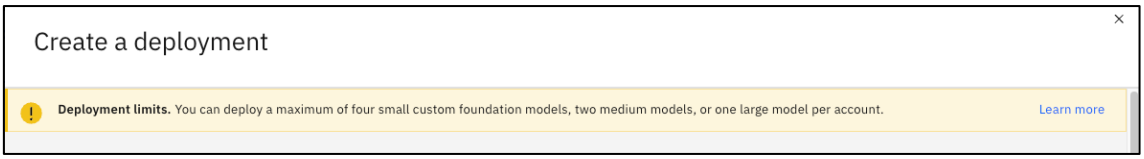


14. Go back to the **Create a deployment** window. Instead of **Create**, you will now see **Reload**. Click **Reload**.

© 2024 IBM Corporation



The message of the top changes to the following:



15. Note that the message says you can create four small, two medium, or one large model. In this exercise, the model you will deploy is a small one.

16. The **Deployment type** is set to **Online** – this is informational and cannot be changed.

17. For **Name** – this is the model name. Enter **CFM-flan-t5-base.**

18. For Description, enter: "**This is a Custom Foundation Model based on flan-t5-base.**"

19. Under **Define configuration**, the **Architecture** is set to **t5**. Leave this as is.

20. Note that there is a cost to model deployment. This means two things:

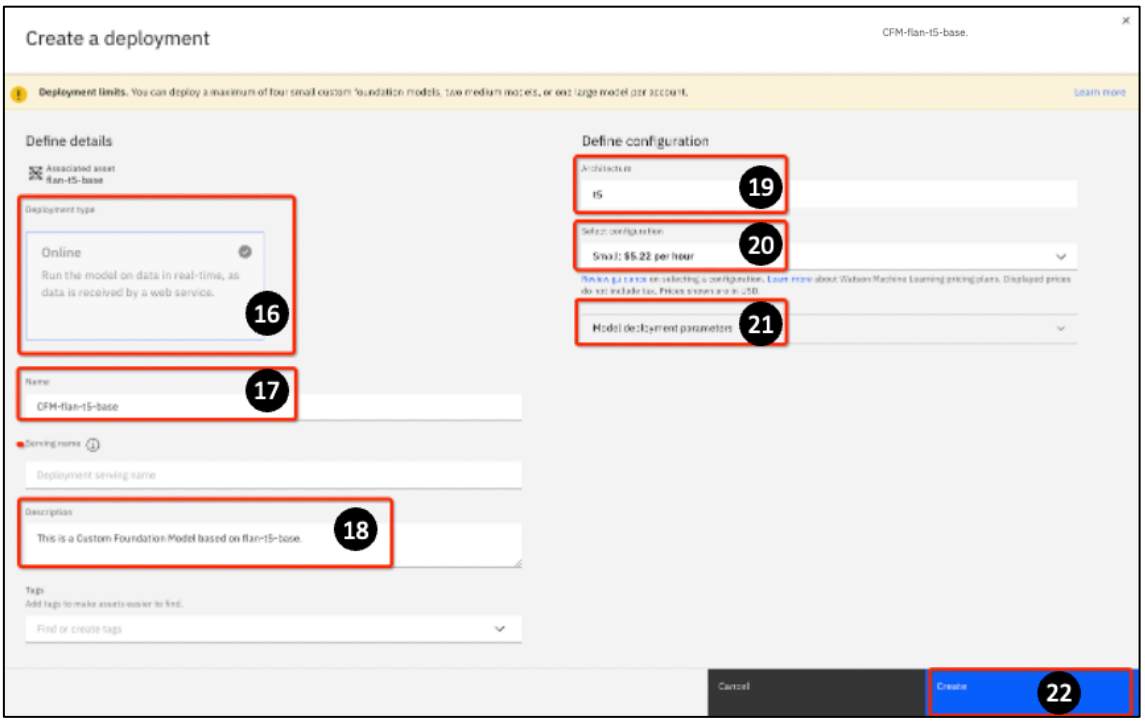
    a. Please be considerate – this exercise uses a small model for a small deployment. That is sufficient for a demo. You should not use a medium or large unless you are in a situation that calls for larger-sized models, for example, if a client demands it in a Proof of Experience exercise.

    b. **You must remove** your deployment **after you have completed the lab** using the instructions in Section 6. Otherwise, these costs will continue to be incurred and charged back to you.

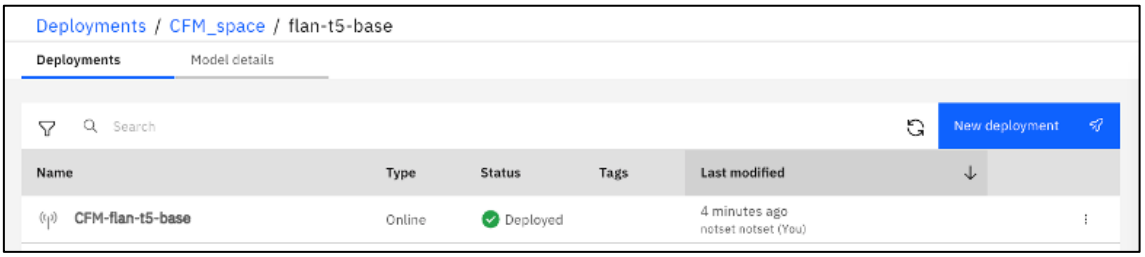
21. You can expand and look at **Model deployment parameters** if so desired.

22. Click **Create**.

© 2024 IBM Corporation



The deployment will take some time to complete. In this example, it took 4 minutes.



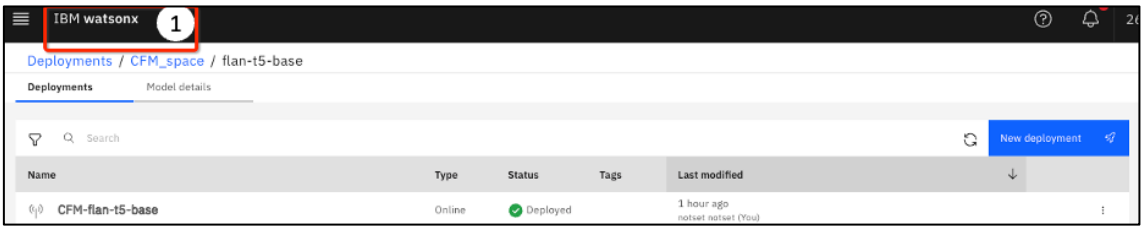
This model is now ready to be used.

© 2024 IBM Corporation

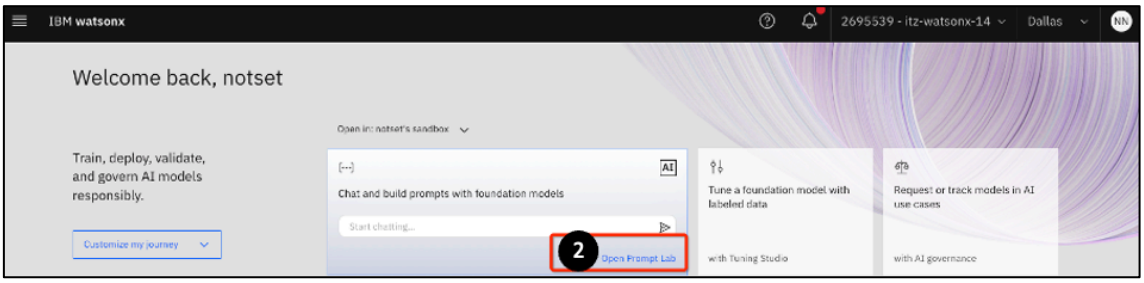
5.9 Using the Custom Foundation Model

Once a model is deployed, you can use it like other models available in watsonx.ai.

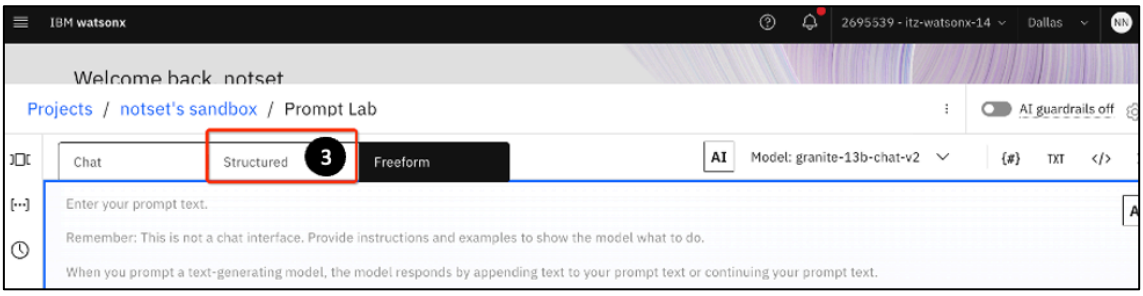
1. Open up the watsonx.ai console. If you have not closed the browser tab from Step 22 of the last section (Section 5.8), click **IBM watsonx**.



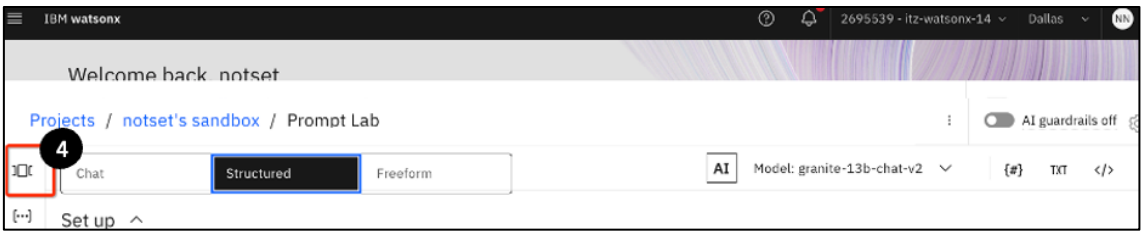
2. The **watsonx.ai** console opens. Click **Open Prompt Lab**.



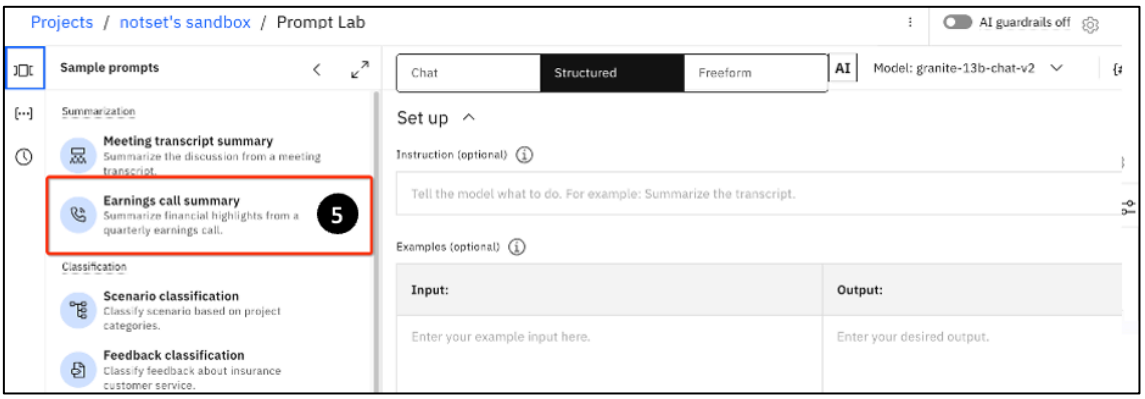
3. Click the **Structured** tab if watsonx.ai does not open in it.



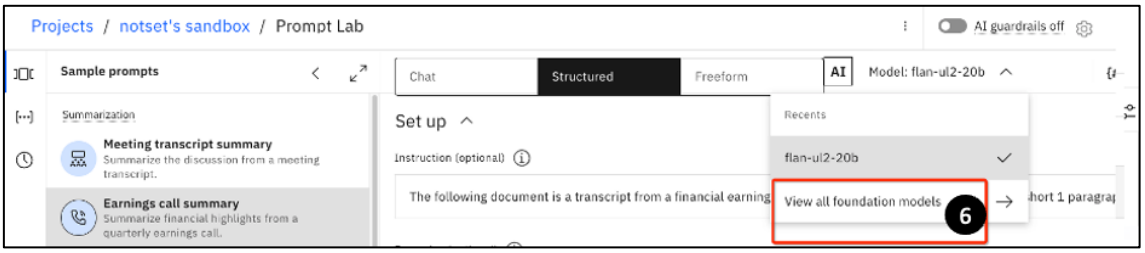
4. Click the **Sample prompts** icon.



5. Click **Earnings call summary** from the list of **Sample prompts** slide-out.

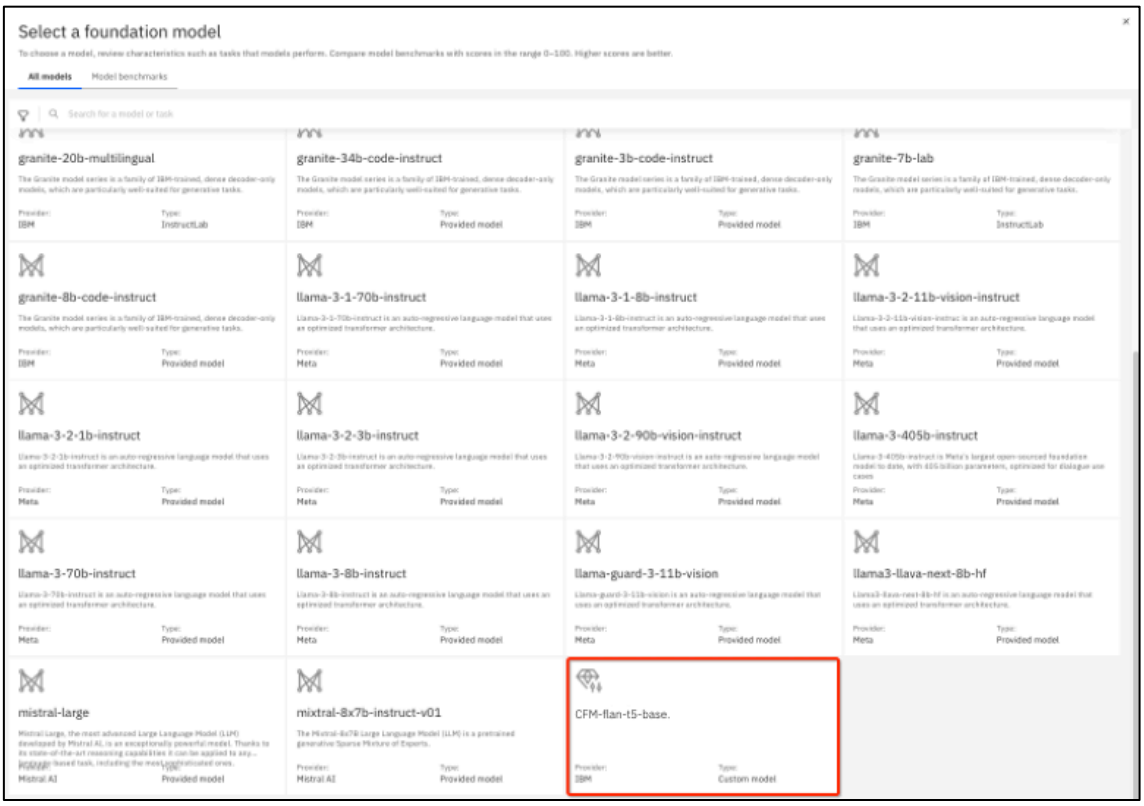


6. The example earnings call summary is loaded. Click on the default **flan-ul2-20b** model, then click **View all foundation models**.



7. Scroll down to find and click on your deployed model: **CFM-flan-t5-base.**

© 2024 IBM Corporation



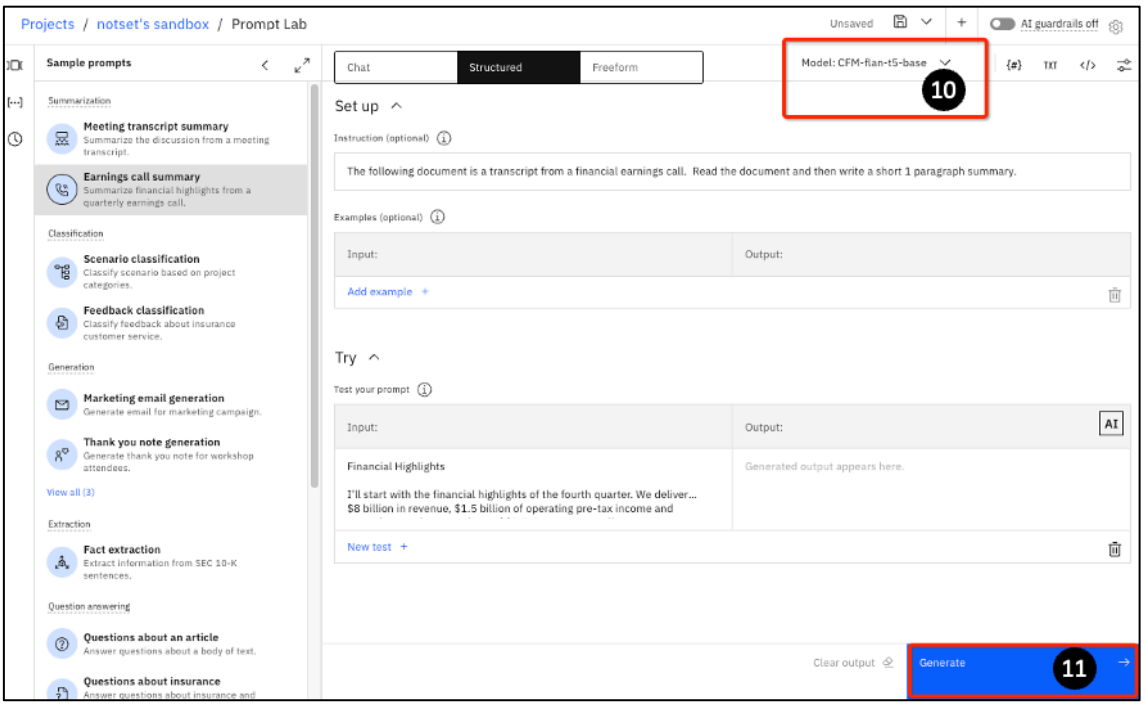
8. The **CFM-flan-t5-base** model card opens. Note that the **Description** entered in Step 18 or Section 5.8 appears. You can use the **Description** field to provide more details about the Custom Foundation Model if necessary.

9. Click **Select model**.

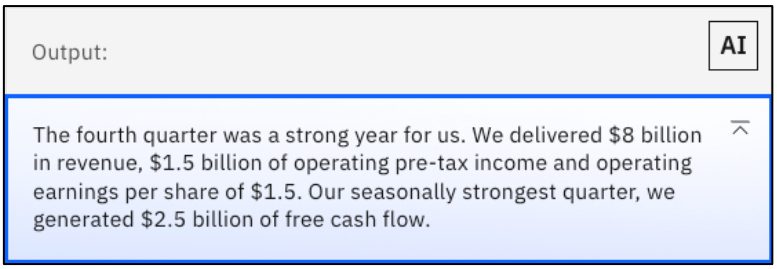
10. The **CFM-flan-t6-base** model is selected for this example.

11. Click **Generate**.

© 2024 IBM Corporation



You will see the following output:



As a side note, this may not be the best summary (the default flan-ul2-20b does better), but this lab was only meant to show that once you have imported a Custom Foundation Model, it can be used just like any other models in the watsonx.ai studio.

© 2024 IBM Corporation

5.9.1 Considerations for Custom Foundation Models

Custom Foundation Model is a powerful tool that allows clients to bring models that they are familiar with, or have already tuned, to the watsonx.ai platform and leverage the various features and capabilities available there. This lab showed the process of importing a model.

Here are some further considerations.

1. Not all models can be imported. Always check the supported architecture link to ensure the model they have in mind is supported.

2. Hugging Face models typically include a model card. However, the current import process does not pull this in. As described in Step 18 of Section 5.8, you can enter information relevant to the model as the **Description** when the model is deployed.

3. Currently, in a watsonx.ai account, you can deploy a maximum of four small models, two medium models, or a single large model (you can, of course, remove deployments before trying a new one if the limit is reached). This is meant for deploying tried and true, known models versus a way to bring in different models to test/try out if they work for your business use cases.

4. There is a cost associated with hosting a Custom Foundation Model, and that depends on the size of the model. See the Hourly billing rates in the custom foundation models section in Watson Machine Learning plans and compute usage.

5. There are some current limitations (for example, you cannot tune a Custom Foundation Model). See Planning to deploy a custom foundation model to find the details.

You must complete the steps outlined in Section 6 as a final task for this lab.

© 2024 IBM Corporation

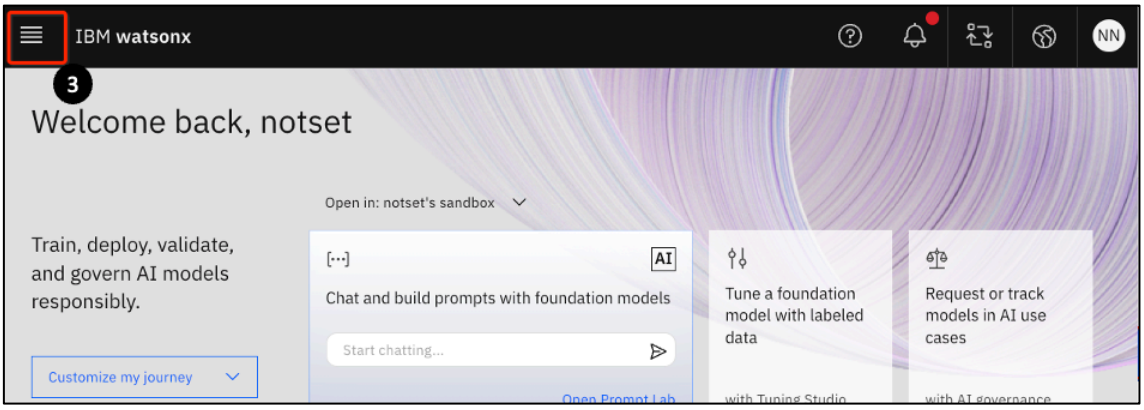
6. Removing a deployment

You must remove your deployment when you are finished with this exercise, or there will be an ongoing cost that will be charged to you.

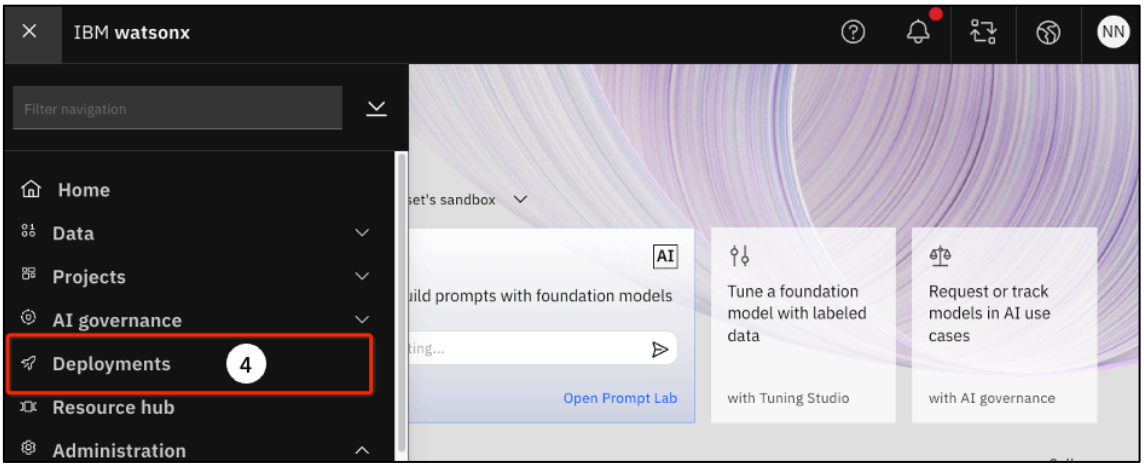
1. If you have logged out of your Student ID account, log back in using the process outlined in Section 5.4.2. If you have not, skip to point 3.

2. Once logged in, bring up the watsonx.ai console using steps 1-10 of Section 5.3.

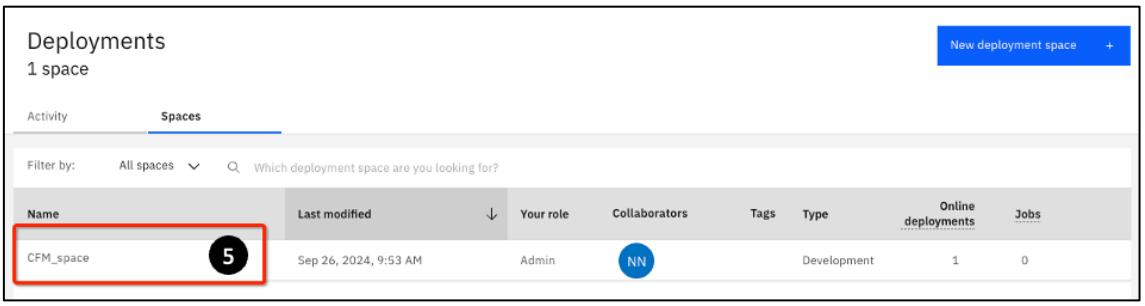
3. Click the main menu on the upper left.



4. Scroll down on the slide-out menu and click **Deployments**.

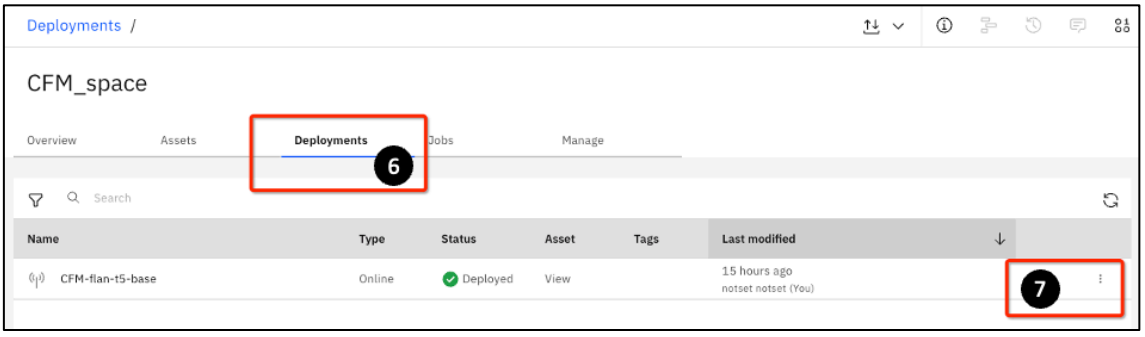


5. The **Deployments** page opens. Click **CFM_space**.

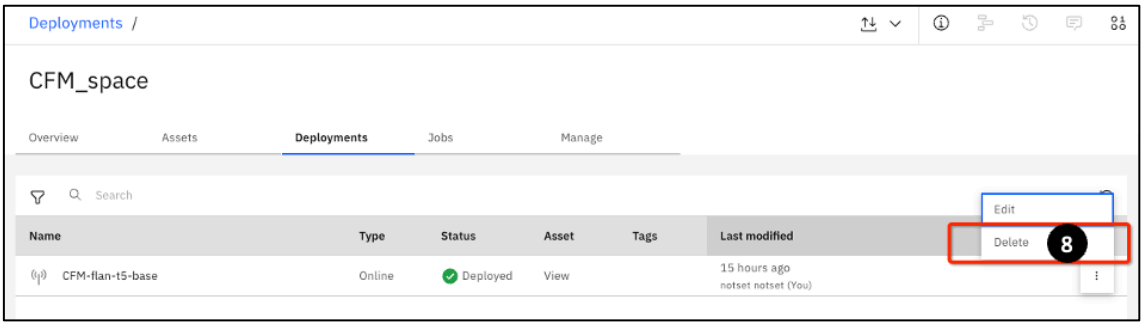


6. The **CFM_space** page opens. Click the **Deployments** tab if it did not open to it.

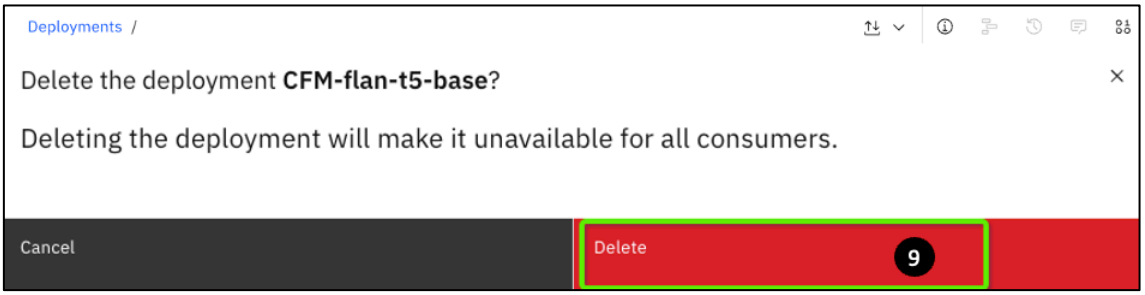
7. Click on the three vertical dots corresponding to **CFM-flan-t5-base** deployment space.



8. Click **Delete**.



9. Click **Delete**.



© 2024 IBM Corporation

Appendix A. Revision History

Date	Changes
October 2024	Original version.

© 2024 IBM Corporation