Retrieval Augmented Generation (RAG) with IBM watsonx.data and watsonx.ai leveraging Milvus vector database

Hands-on lab guide

Pratik Sinha (Pratik.sinha1@ibm.com)
AI Engineer, Ecosystem Engineering Lab, Client Engineering

Contents

1 About this lab

This lab exhibits how watsonx.data and watsonx.ai can be integrated using Milvus vector database. This lab showcases a step-by-step RAG (Retrieval Augmented Generation) approach to perform Q&A task on watsonx documents.

Note: The lab requires that you have a watsonx.data and watsonx.ai environment up and running already. See the Prerequisites & Getting Started section for details.

2 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). Business Partners can request help at the Partner Plus Support website. You can also reach out to me directly through slack or email (slack id- @Pratik Sinha | email- Pratik.Sinha1@ibm.com).

Techzone environment: If you are encountering issues regarding the Techzone environment being used in this lab, including the inability to provision an environment, please see the Techzone Help page.

watsonx.data: Assistance with the watsonx.data product itself is available in the #watsonx-datalakehouse-discussion-open-to-all-ibmers Slack channel (IBMers only). Additionally, please refer to the watsonx.data documentation as needed (SaaS, software).

3 Prerequisites & getting started

- watsonx.data Environment: This lab requires that you already have a provisioned IBM watsonx.data environment. IBMers and Business Partners can easily provision a no-cost environment from IBM Technology Zone (TechZone). The IBM watsonx.data Developer Base Image has to be used for this lab. For detailed instructions on how to provision and use this environment, see the “Prerequisites & Getting Started” section of this other IBM watsonx.data hands-on lab.

  Note: The screenshots used in this lab guide are from the Developer Edition but should be like what is seen in the other deployment options. You are expected to provision the **IBM watsonx.data Development Lab – 2.1.1 GA environment**.

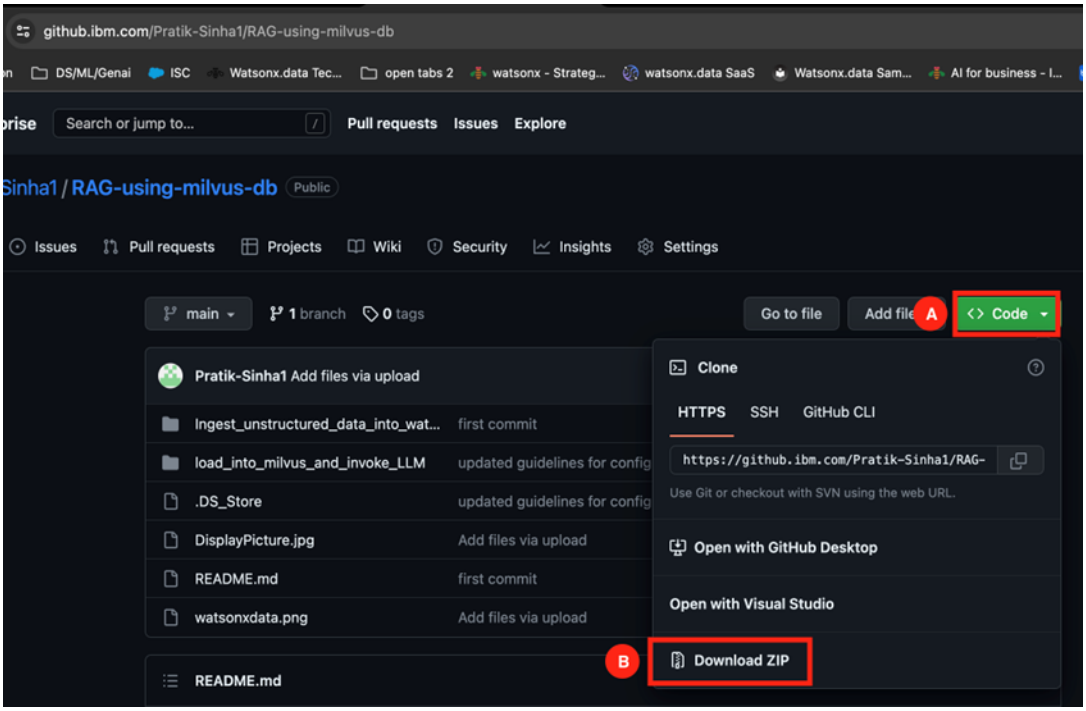
- watsonx.ai Environment: This lab requires that you already have a provisioned IBM watsonx.ai environment. IBMers and Business Partners can easily provision a no-cost environment from TechZone. The IBM watsonx.ai Developer Base Image must be used for this lab. For detailed instructions on how to provision and use this environment, see the “Prerequisites & Getting Started” section of this other IBM watsonx.ai hands-on lab.

- TechZone IBM cloud account: After the watsonx.ai instance has been provisioned; you will be invited to join a TechZone account.

- GitHub link- RAG-using-milvus-db (You will be downloading and using the Jupyter notebooks from this GitHub repository)

4 Download the lab files to your system

1. Go to the GitHub page: [RAG-using-milvus-db.]
2. Click <> Code (A).
3. Click **Download ZIP (B)** to download the file to your system.



4. Unzip the file by double clicking it in your system (on Mac) or by right clicking the file and clicking Extract All on Windows).

   You should see the RAG-using-milvus-db folder is now present in your system.

5 Running the notebooks

This lab requires you to run two notebooks. You will also need to update a configuration file to configure the environment as mentioned in the below steps. You will execute the lab in two steps each using a notebook to showcase the functionality.

Notebook 1 - Ingesting unstructured data into watsonx.data.

Notebook 2 - Create vector embeddings, ingest them into Milvus vector database, and finally invoke Large Language Model (LLM) from watsonx.ai to apply to your generative AI Q&A use-case.

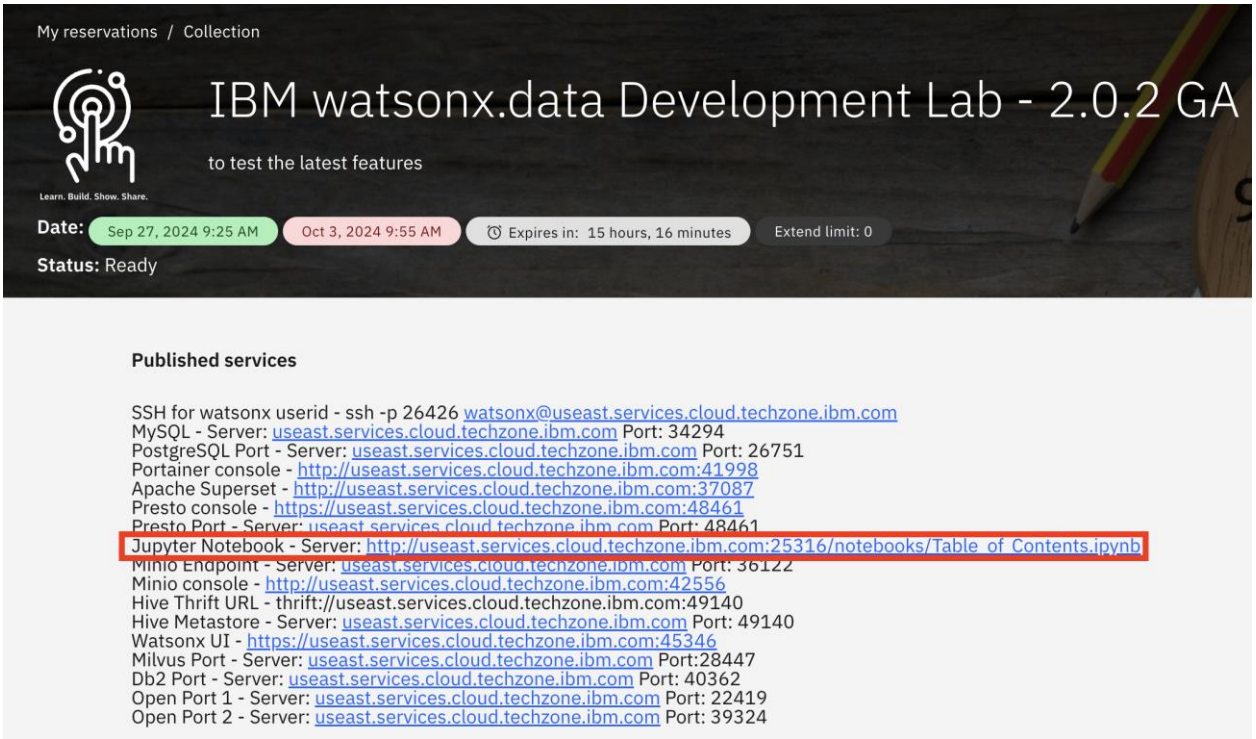
6 Working with notebook 1

6.1 Import files into watsonx.data jupyter server

In your lab folder which you unzipped earlier on your local system, you should see two folders. For this section, you would need to import the folder- Ingest_unstructured_data_into_watsonx.data into watsonx.data Jupyter notebook file storage.

1. Go to your TechZone reservation page – https://techzone.ibm.com/my/reservations and click the **IBM watsonx.data Development Lab 2.1.1 GA** tile. It will open the list of Published services associated with your instance and should look like the image below.

   Note: The images/screenshots in this guide might be from the previous environments. You are expected to work with **IBM watsonx.data Development Lab 2.1.1 GA** environment.

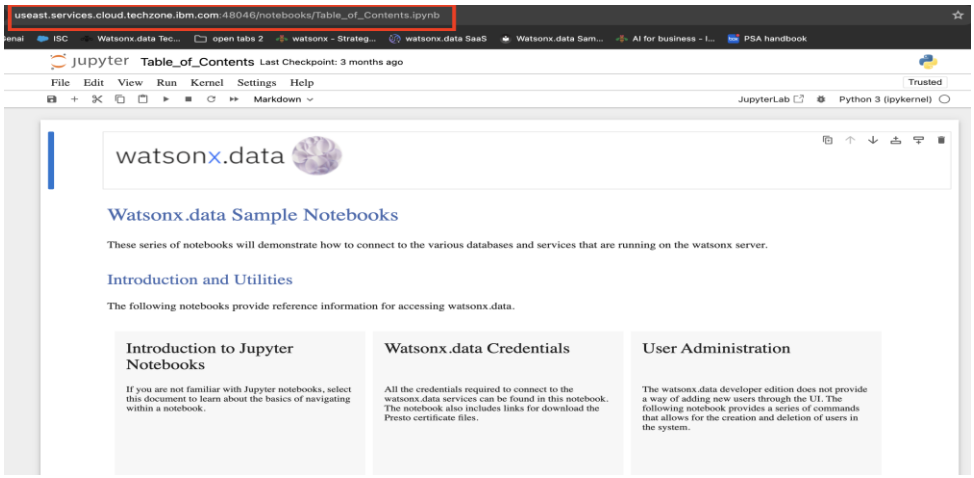


2. Click the **Jupyter Notebook – Server** link and it will open in a new tab.

3. Fill in the password - **watsonx.data** and click **log in**.

   Go to the URL field at the top of your browser window. Note the URL, which should look like the following:
   http://useast.services.cloud.techzone.ibm.com:25316/notebooks/Table_of_Contents.ipynb.

   Note: Your port number will be different from this example above and from the screenshots.

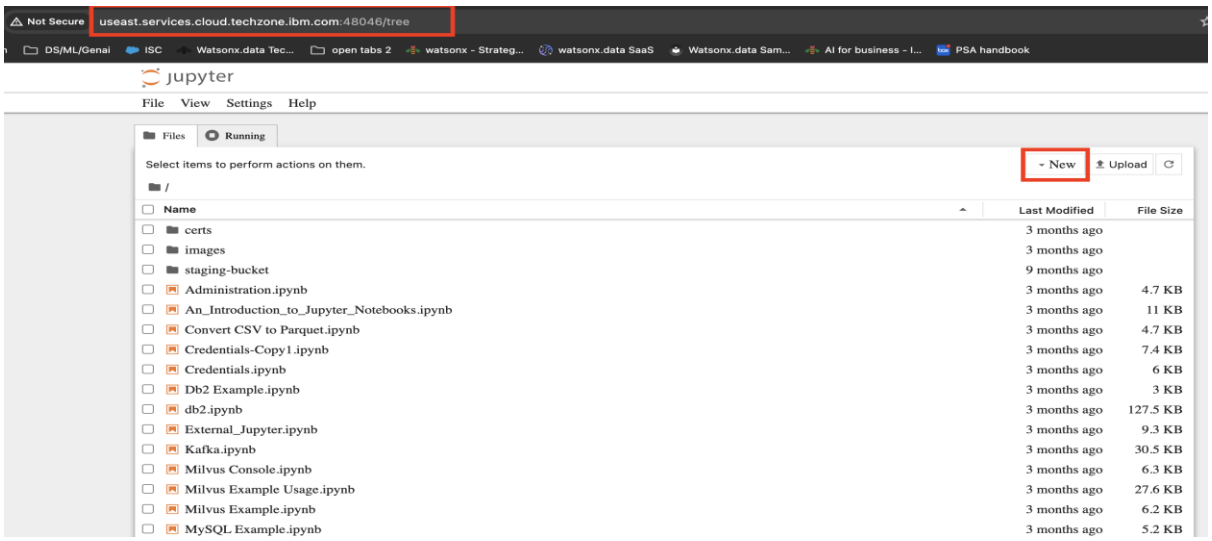


4. Open the Jupyter notebooks file storage by replacing the text in the URL “notebooks/Table_of_Contents.ipynb” with “tree”as in the example below.

   Initial URL:
   http://useast.services.cloud.techzone.ibm.com:25316/notebooks/Table_of_Contents.ipynb

   Updated URL:
   http://useast.services.cloud.techzone.ibm.com:25316/tree

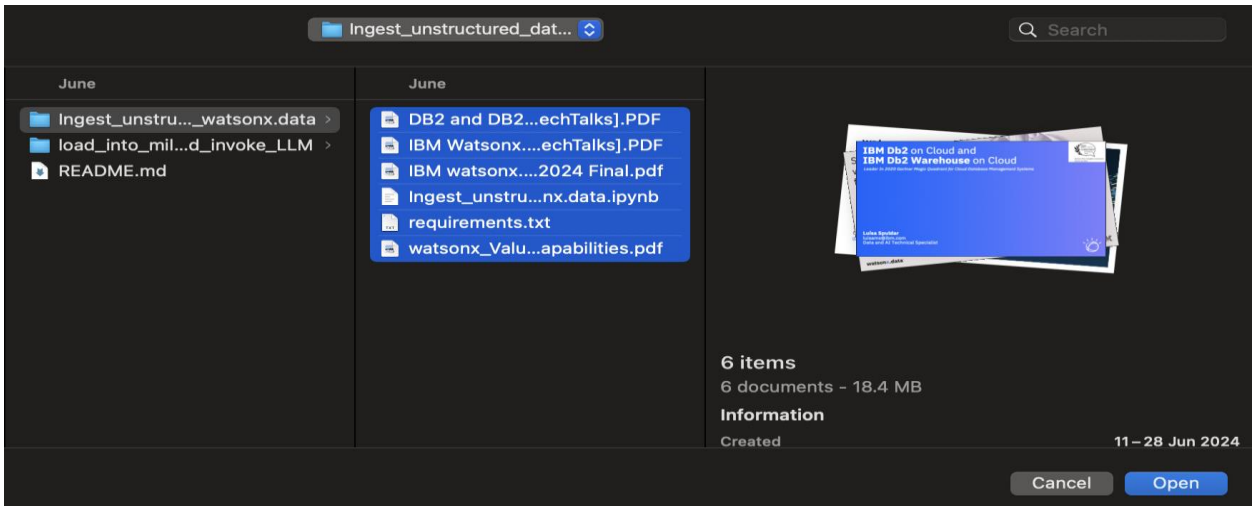
   Once done, you will see a list of folders and notebooks.



5. On the top right side, click **New** and select **New Folder**. Give the same name as the folder that will be imported - **Ingest_unstructured_data_into_watsonx.data**.

6. Once created, navigate to that folder by double clicking it and click the **Upload** button.

7. Select all the files present in the **Ingest_unstructured_data_into_watsonx.data** folder in your system and click **Open**.



You should now see the files being uploaded into the server.

6.2 Open and run the jupyter notebook

Double-click on the **Ingest_unstructured_data_into_watsonx.data.ipynb** notebook for the next steps. Follow the instructions and details provided in the notebook. Once you've completed running the notebook, return to this lab guide.

6.3 Summary of notebook 1

In this notebook, you learned how to:

- Run a python notebook
- Read unstructured data from PDF documents
- Connect to watsonx.data and run SQL queries
- Create schemas and tables in watsonx.data
- Chunk the data using RecursiveCharacterTextSplitter algorithm
- Ingest data into a watsonx.data table

7 Working with notebook 2

In this notebook, you will vectorize/embed the data and load it into the Milvus vector database. Once the embeddings have been loaded into Milvus, you will then query the vector database, extract the relevant data and send them to the LLM inside watsonx.ai to generate results.

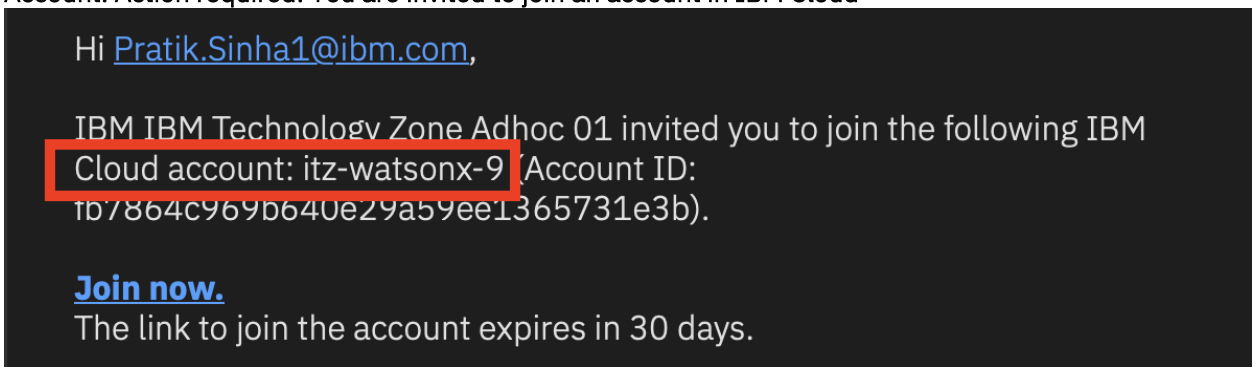
7.1 Setup watsonx.ai project

In this section, you will create a project inside watsonx.ai.

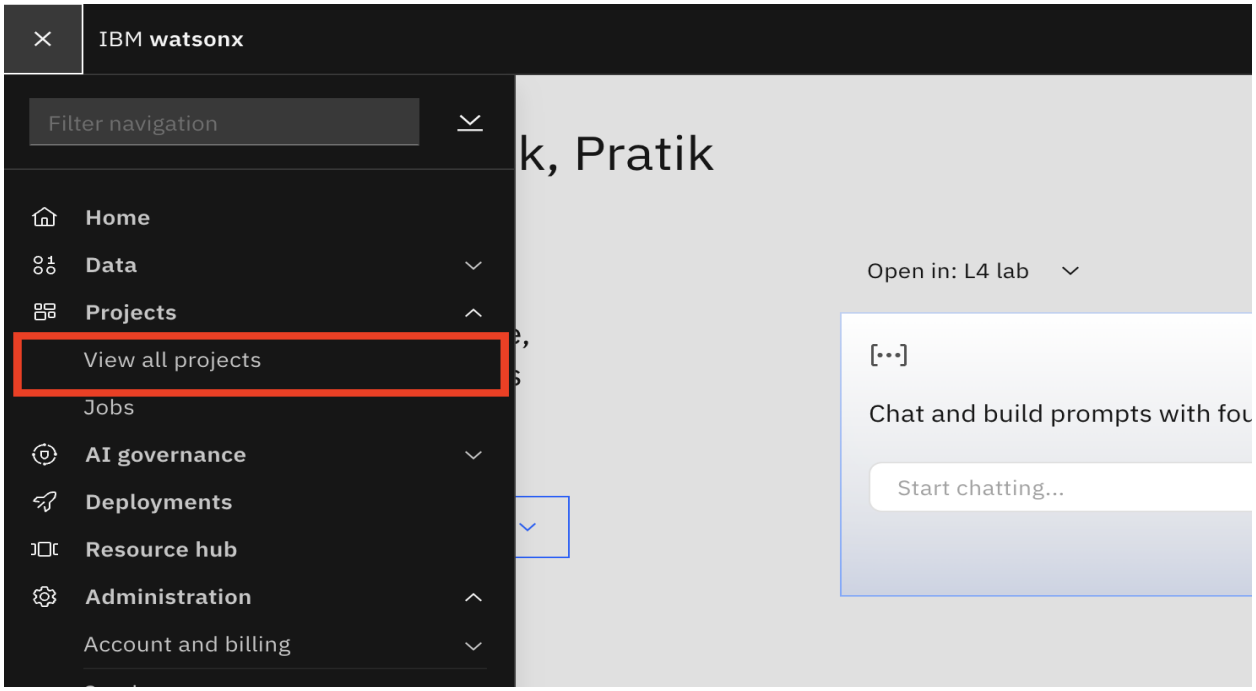
1. Navigate to the watsonx.ai homepage.

2. Ensure the cloud account that you are working is the current provisioned one. You can check this through the top header of the platform.

Make sure this is the same cloud account which is mentioned in the e-mail you received with subject- Account: Action required: You are invited to join an account in IBM Cloud

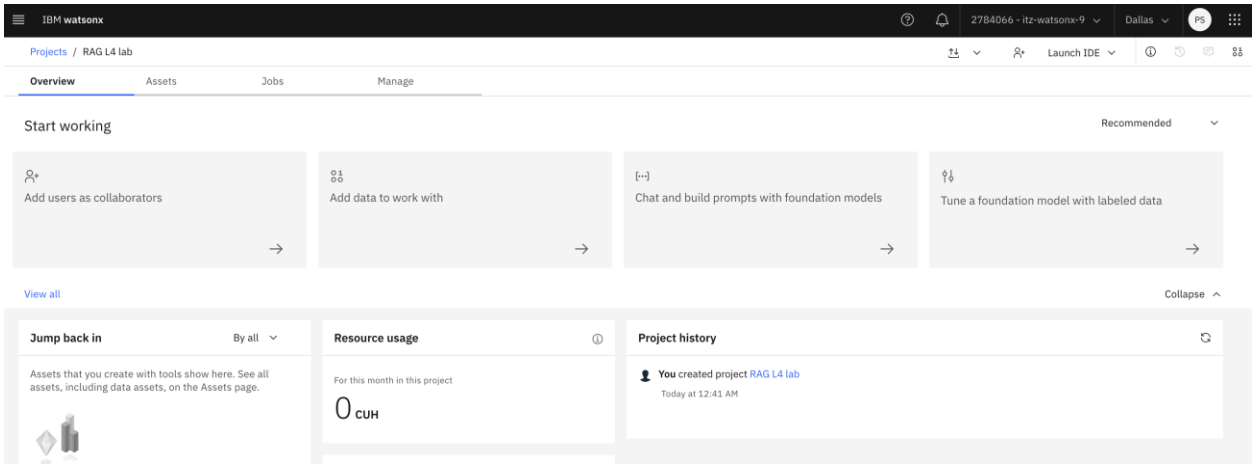


3. Click the hamburger menu on top left of the main menu and select View all projects.



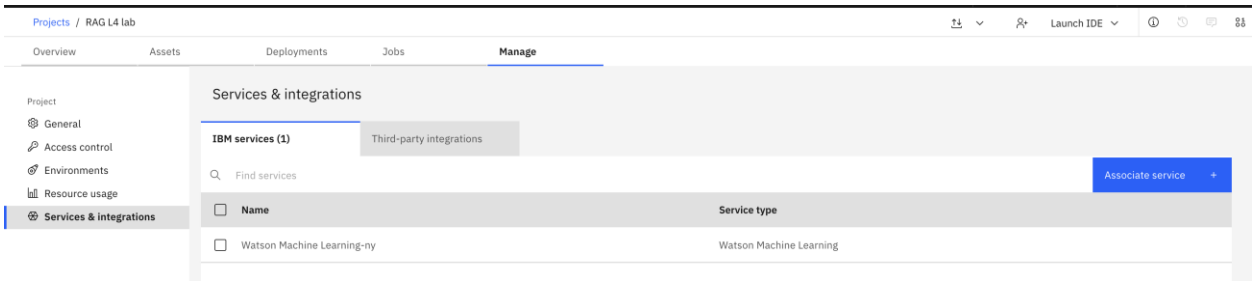
4. Click New Project +.

5. Enter the name of the project as RAG L4 lab and click Create. Once created, you should see a screen like this.



6. Click **Manage.**
7. Click **Services & integrations.**
8. You should see a Machine Learning service associated to the project.

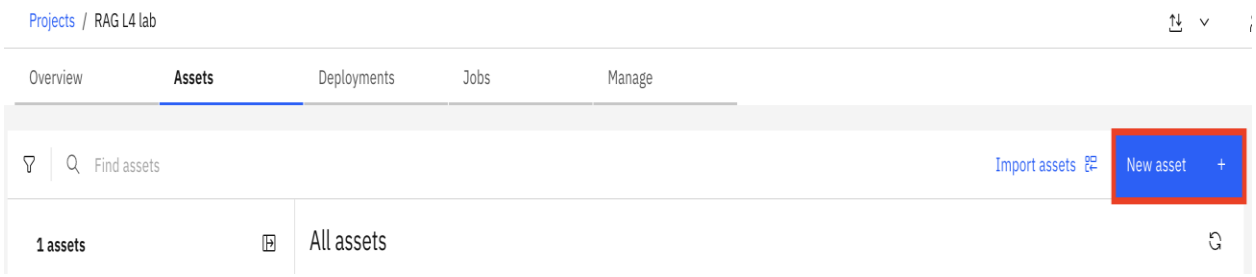
   If not, click on **Associate Service +** blue button --> check mark the Machine Learning service available--> click **Associate**.



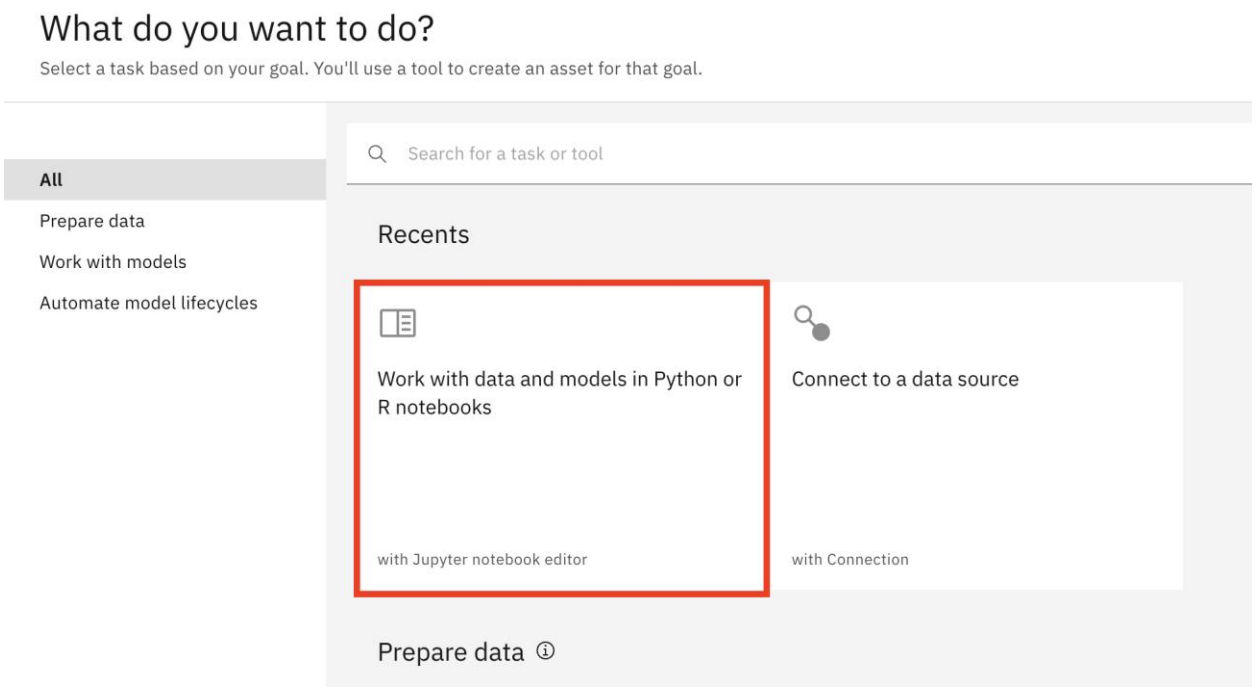
7.2 Import notebook 2 into the watsonx.ai project

In this step, you will import the Jupyter notebook into the watsonx.ai project. To do so, follow the below steps:

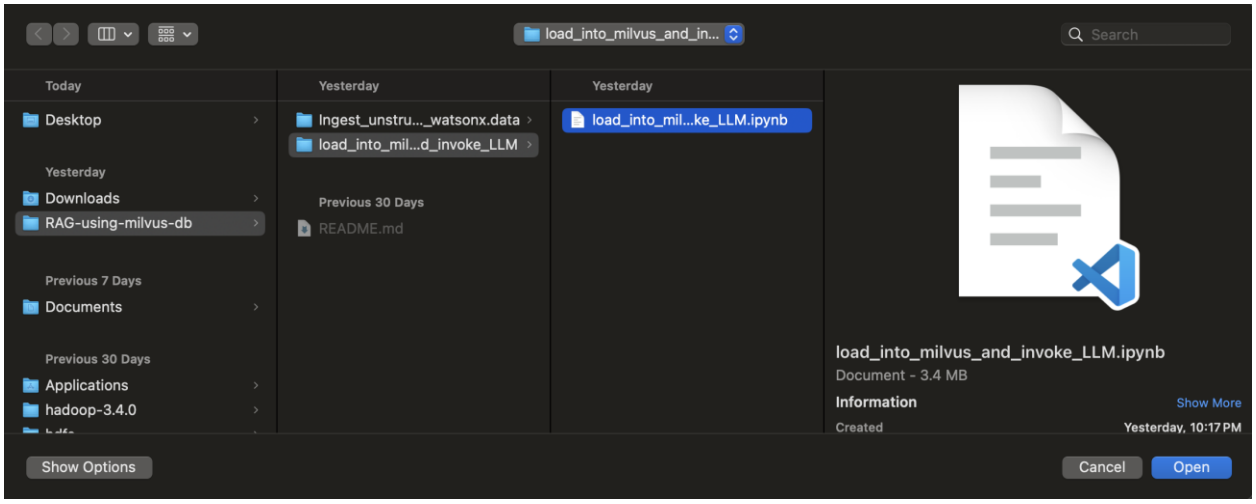
1. Click **Assets** on the project page (Refer previous screenshot).
2. Click **New asset +** button (highlighted in blue).



3. Select **Work with data and models in Python or R notebooks** tile.

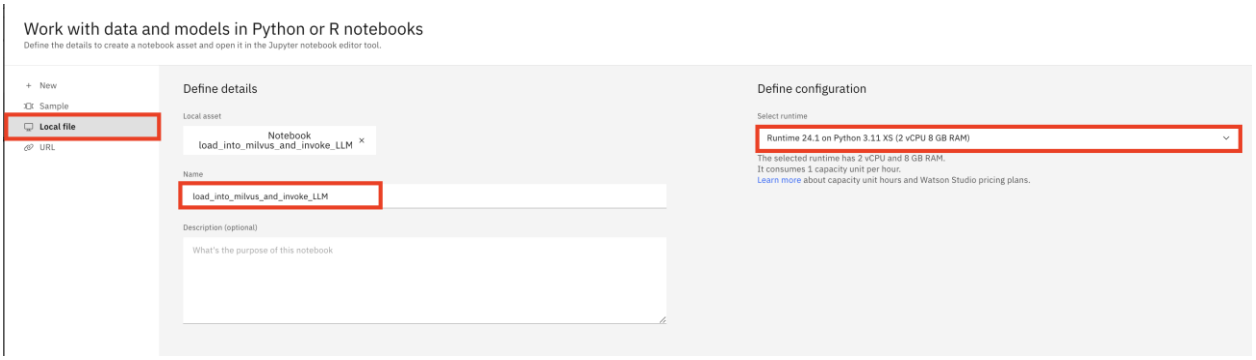


4. Click **Local file**.
5. Click **Browse**.
6. Navigate to the **RAG-using-milvus-db** folder in your local system.
7. Open the **load_into_milvus_and_invoke_LLM** subfolder.
8. Select **load_into_milvus_and_invoke_LLM.ipynb** file.
9. Click **Open** (refer screenshot below).

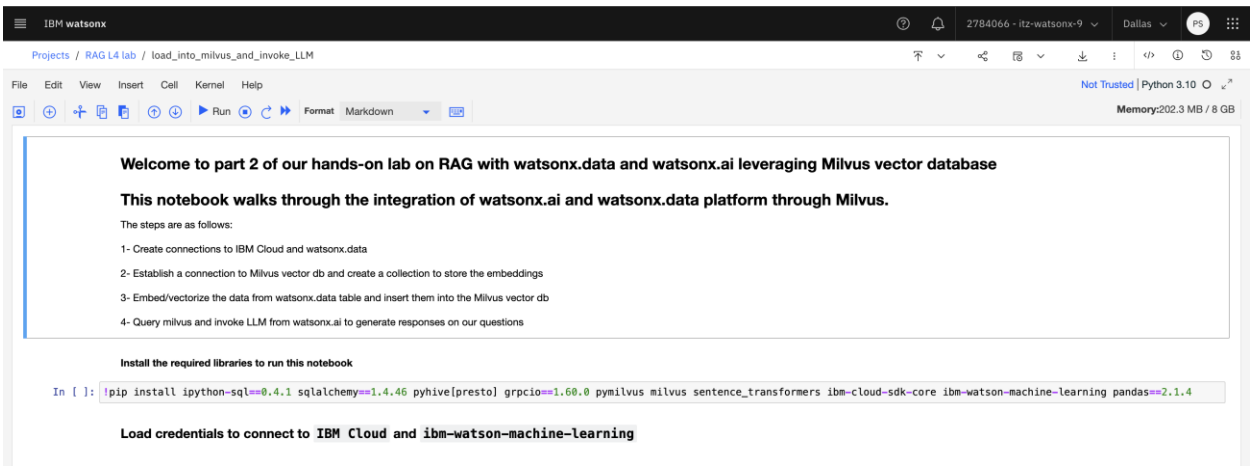


10. Select the **Runtime 24.1 on Python 3.11 XS (2 vCPU 8 GB RAM)** in Select runtime drop-down.

11. Enter the Name- **load_into_milvus_and_invoke_LLM** and click **Create**.



Once created, your Jupyter notebook should be up and ready.



Head over to the notebook now and follow the instructions mentioned to proceed with the lab.

7.3 Summary of notebook 2

In this notebook, you learned how to:

- Create a project in watsonx.ai
- Create API keys and connect to the IBM Cloud
- Establish a connection to watsonx.data and the Milvus vector database
- Create a collection in Milvus
- Vectorize the data using the all-MiniLM-L6-v2 embedding model
- Load vector embeddings to the Milvus vector database
- Query Milvus to perform similarity search and retrieve results
- Invoke an LLM from watsonx.ai and generate results for your questions

8 Summary

Congratulations, you have now completed the lab!

In this lab, you learned how to integrate watsonx.data with watsonx.ai using Milvus and create a RAG Q&A use case.